COMP5046
Natural Language Processing

Lecture 10: Attention and Question Answering (Reading Comprehension)

Semester 1, 2020
School of Computer Science
The University of Sydney, Australia

THE UNIVERSITY OF SYDNEY

Dr Caren Han
Caren.Han@sydney.edu.au

0 LECTURE PLAN

Lecture 10: Attention and Question Answering (Reading Comprehension)

Recap Lecture 9 and NER with Bi-LSTM-CRF

1. Question Answering
2. Knowledge-based Question Answering
3. IR-based Question Answering (Reading Comprehension)
4. Attention
5. Reading Comprehension with Attention
6. Visual Question Answering

Recap Lecture 9 (NER) and Bi-LSTM CRF

Named Entity Recognition

The goal: predicting named entity mentions in unstructured text into pre-defined categories *such as the person names, organizations, locations*

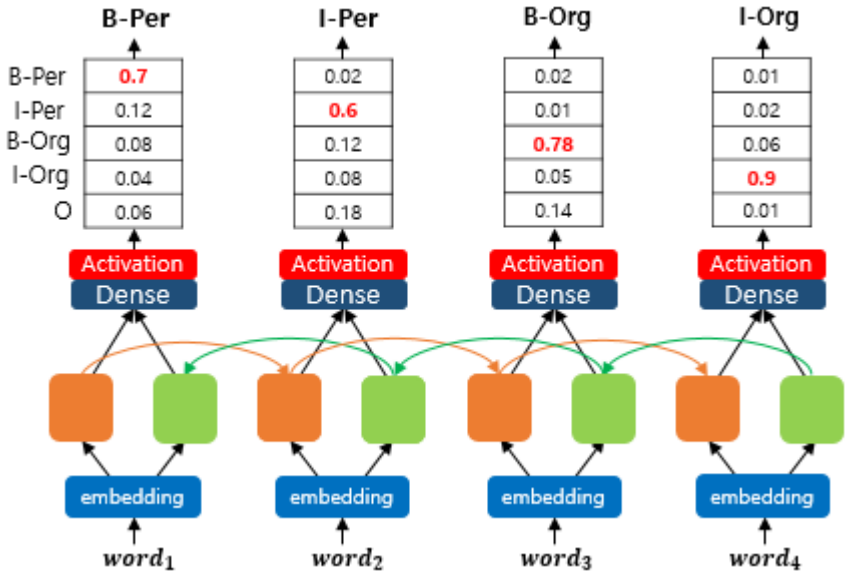
Upenn CogComp-NLP *http://macniece.seas.upenn.edu:4004/*

	Caren	Soyeon	Han	is	working	at	Google	at	Sydney,	Australia
gold	PER	PER	PER	O	O	O	ORG	O	LOC	LOC
predicted	O	O	O	O	O	O	ORG	O	LOC	LOC

Recap Lecture 9 (NER) and Bi-LSTM CRF

Named Entity Recognition with Bi-LSTM

We can easily apply Bi-LSTM (N to N Seq2Seq) Model to predict Named Entities



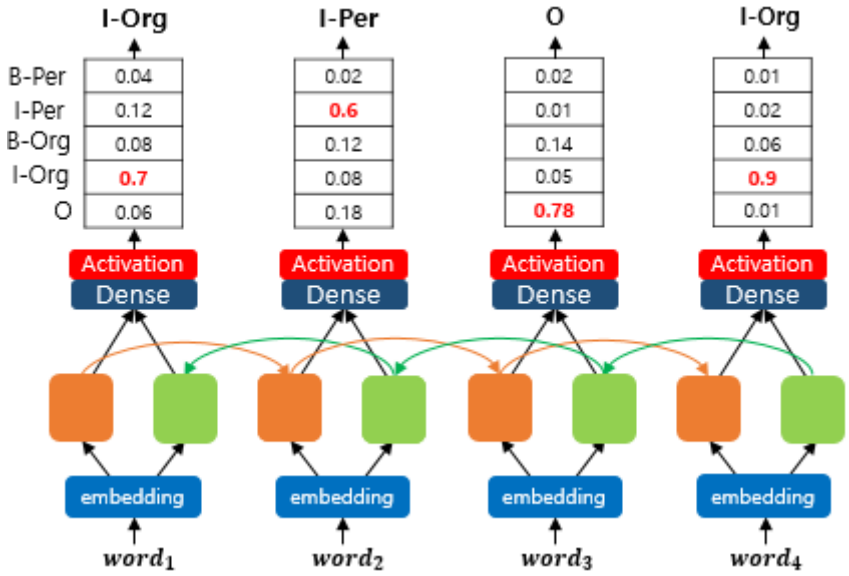
Recap Lecture 9 (NER) and Bi-LSTM CRF

Named Entity Recognition with Bi-LSTM

We can easily apply Bi-LSTM (N to N Seq2Seq) Model to predict Named Entities

*The model clearly contains incorrect predictions.

'I' cannot appear in the label of the first word. I-Per can only appear after B-Per. I-Org can also appear only after B-Org.



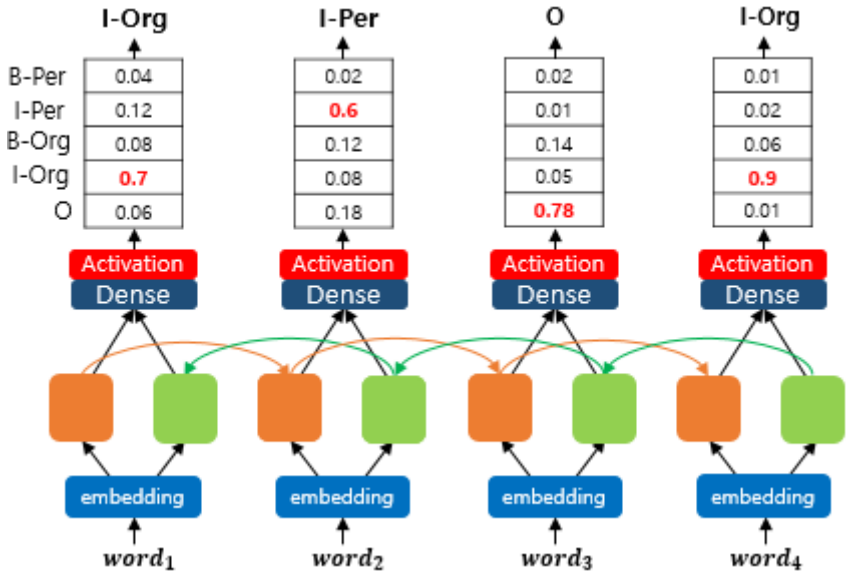
Recap Lecture 9 (NER) and Bi-LSTM CRF

Named Entity Recognition with Bi-LSTM

We can easily apply Bi-LSTM (N to N Seq2Seq) Model to predict Named Entities

*The model clearly contains incorrect predictions.

'I' cannot appear in the label of the first word. I-Per can only appear after B-Per. I-Org can also appear only after B-Org.

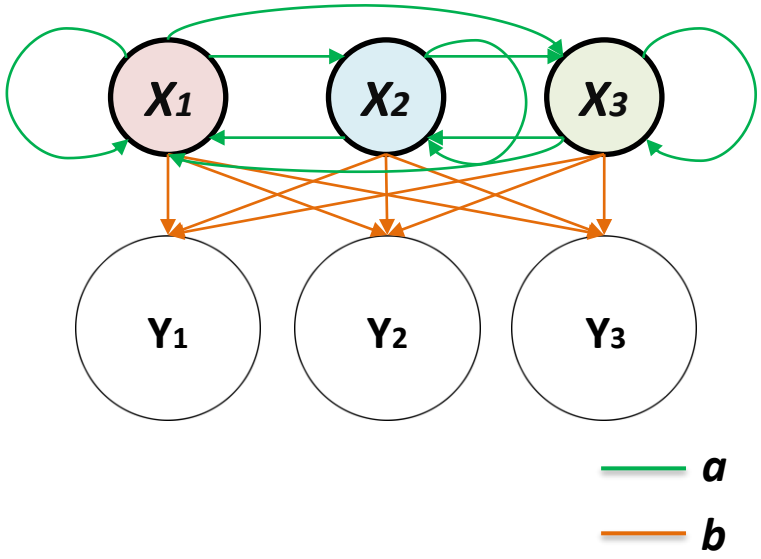


What if we teach the dependency between predicted entity names

Recap Lecture 9 (NER) and Bi-LSTM CRF

Wait? What about HMM?

Hidden Markov Models (HMMs) are a class of probabilistic graphical model that allow us to predict a sequence of unknown (hidden) variables from a set of observed variables.



hidden

x	states
y	possible observations
a	state transition probabilities
b	output probabilities

- States are ***hidden***
- ***Observable outcome*** linked to states
- Each state has ***observation probabilities*** to determine the observable event

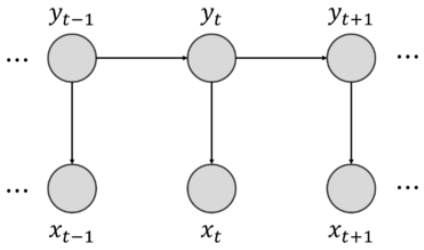
0 Recap Lecture 9 (NER) and Bi-LSTM CRF

THE UNIVERSITY OF SYDNEY

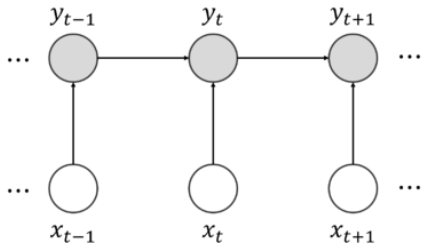
Advanced HMM (MEMM or CRF)

- The CRF model has addressed the labeling bias issue and eliminated unreasonable hypotheses in HMM.
- MEMM adopts local variance normalization while CRF adopts global variance normalization.

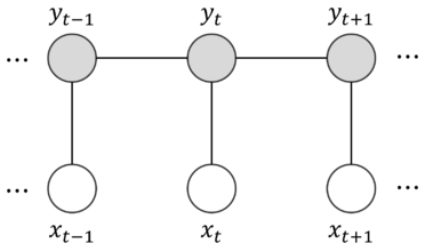
HMM



Maximum-entropy Markov model (MEMM)



Conditional random field (CRF)




6

Part of Speech Tagging

(Lecture 6.pdf / 4slides.pdf)

Suggested Reading:

1) Probabilistic PoS tagging: Speech and Language Processing (Jurafsky and Martin, 2019)

2) Deep Learning LSTM PoS tagger Comparison Study: (Horsmann and Zesch, 2019)

Lecture6 Recording Viewer [Download: Link]

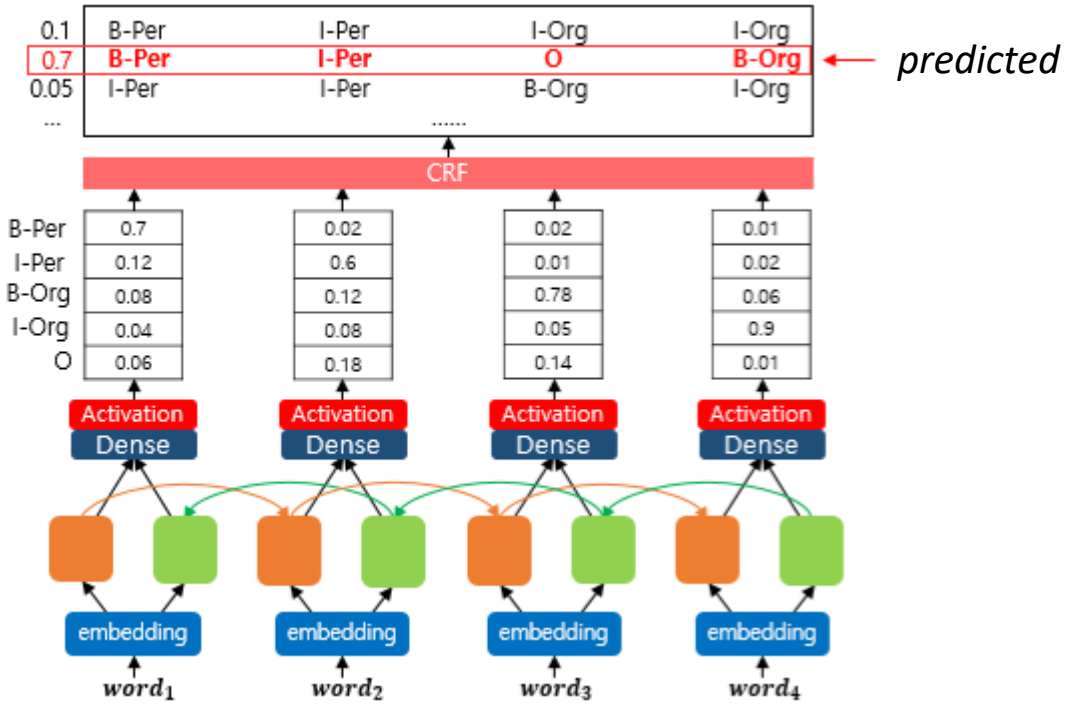
Interesting paper

HMM, MEMM and CRF (Jurafsky and Martin, 2019)

Recap Lecture 9 (NER) and Bi-LSTM CRF

Named Entity Recognition with Bi-LSTM with CRF

What if we put CRF on top of the Bi-LSTM model. By adding a CRF layer, the model can handle the dependency between predicted entity names



0 LECTURE PLAN

Lecture 10: Attention and Question Answering (Reading Comprehension)

Recap Lecture 9 and NER with Bi-LSTM-CRF

1. Question Answering
2. Knowledge-based Question Answering
3. IR-based Question Answering (Reading Comprehension)
4. Attention
5. Reading Comprehension with Attention
6. Visual Question Answering

Question Answering

Question Answering

Question answering (QA) *is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that* ***automatically answer questions posed by humans in a natural language****.*

Different types of questions:

General questions*, with Yes/No answers*

- *e.g. Are you a student?*

Wh- Questions*, start with: who, what, where, when, why, how, how many*

- *e.g. When did you get to this lecture?*
- *e.g. What is the weather like in London?*

Question Answering

Question Answering

Question answering (QA) *is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that* ***automatically answer questions posed by humans in a natural language****.*

Different types of questions:

Choice Questions*, where you have some options inside the question*

Factoid questions*, where the complete answer can be found inside a text. The answer to such questions consist of one or several words that go one after another*

1 Question Answering

THE UNIVERSITY OF SYDNEY

Question

Three Questions for building a QA System

- What do the answers look like?
- Where can I get the answers from?
- What does my training data look like?

Question Answering Research Areas

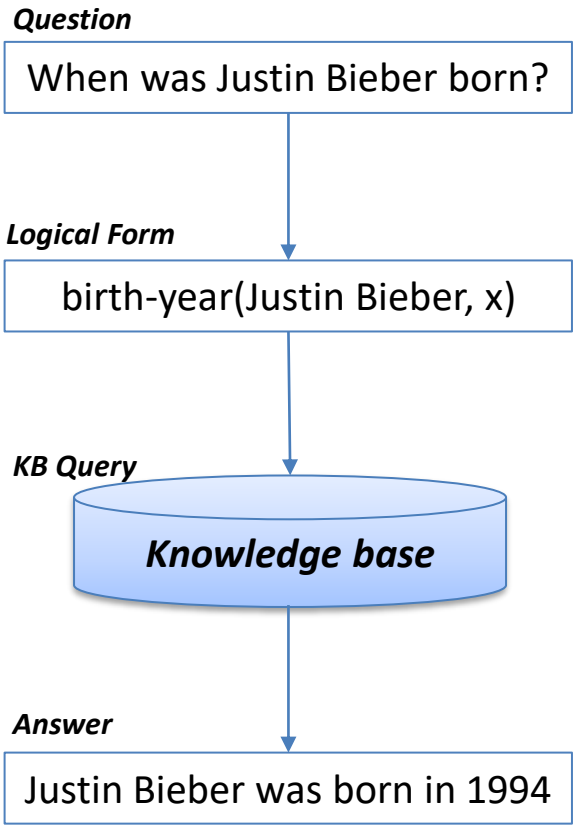
Research Areas in Question Answering

Research Area	Details
Knowledge-based QA (Semantic Parsing)	• Answer is a logical form, possible executed against a Knowledge Base • Context is a Knowledge Base
Information Retrieval-based QA • Answer sentence selection • Reading Comprehension	• Answer is a document, paragraph, sentence • Context is a corpus of documents or a specific document
Visual QA	• Answer is simple and factual • Context is one/multiple image(s)
Library Reference	• Answer is another question • Context is the structured knowledge available in the library and the librarians view of it.

Knowledge-based Question Answering

Semantic Parsing

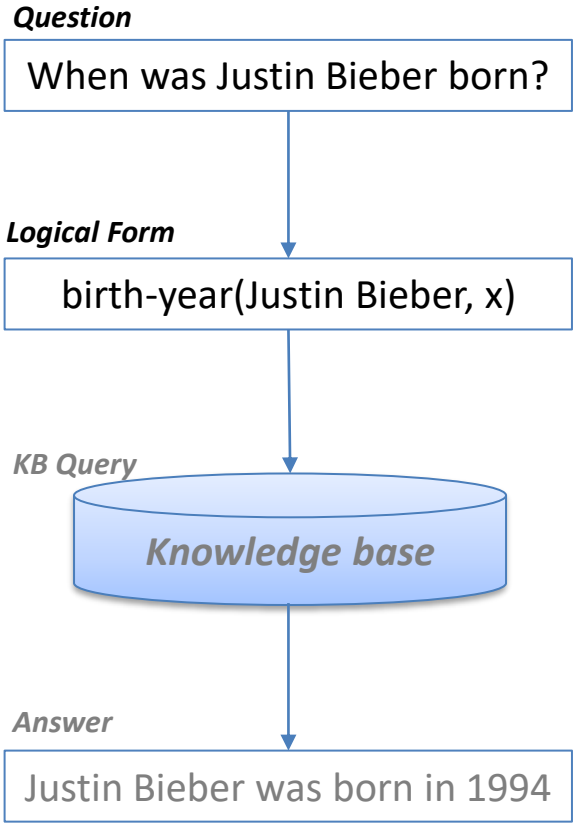
Answering a natural language question by mapping it to a query over a structured database (formal representation of its meaning).



Knowledge-based Question Answering

Semantic Parsing

Answering a natural language question by mapping it to a query over a structured database (formal representation of its meaning).



Mapping from a text string to any logical form

Question	Logical Form
When was Justin Bieber born?	birth-year(Justin Bieber, x)
What is the largest state?	argmax(λx.state(x),λx.size(x))

How to map? Map either to some version of predicate calculus or a query language like SQL or SPARQL
https://query.wikidata.org/

Use expensive supervised data?

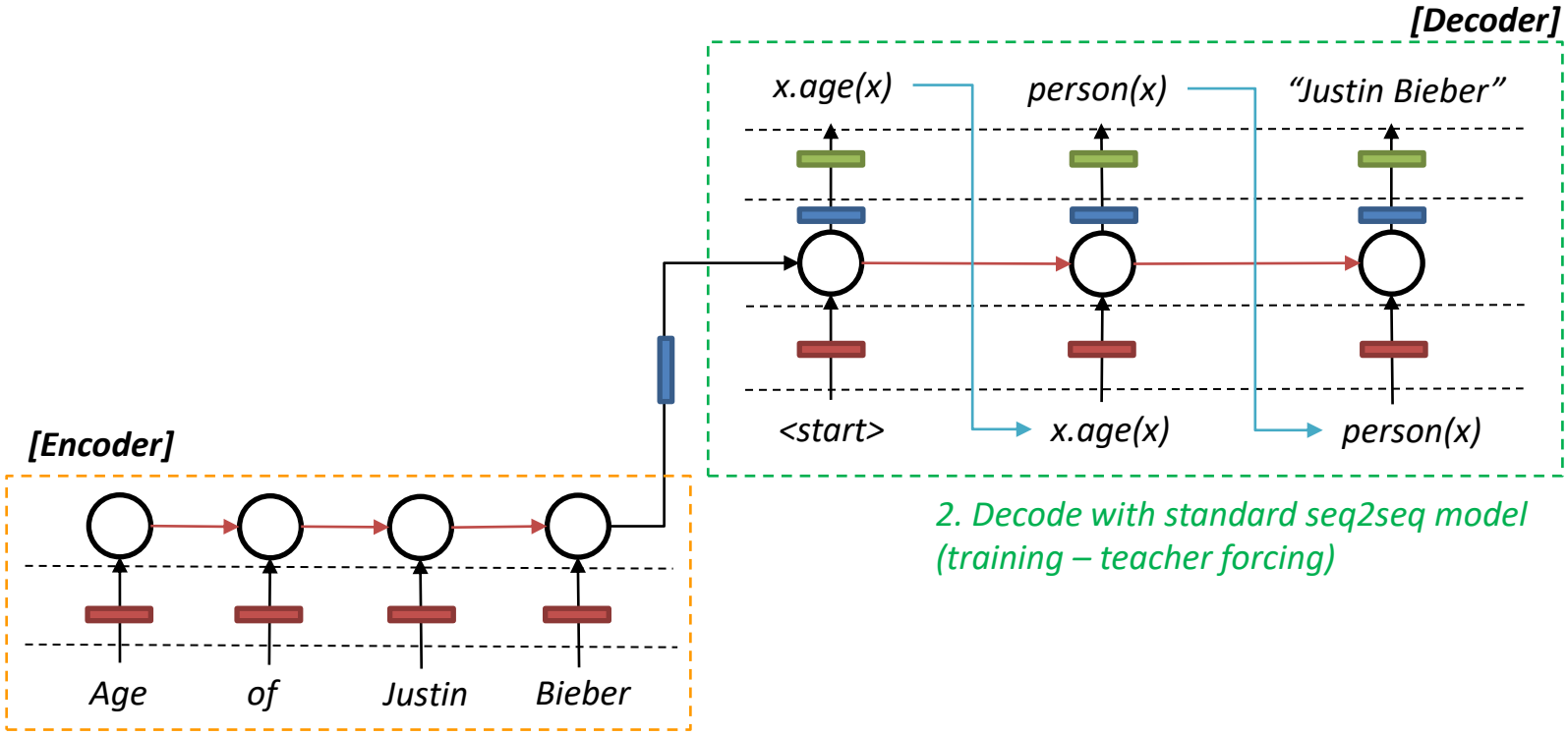
Require experts for the manual annotation process….

Knowledge-based Question Answering

Seq2Seq model for semantic parser

How to transfer the text to the logical form?

A basic deep learning approach to semantic parsing



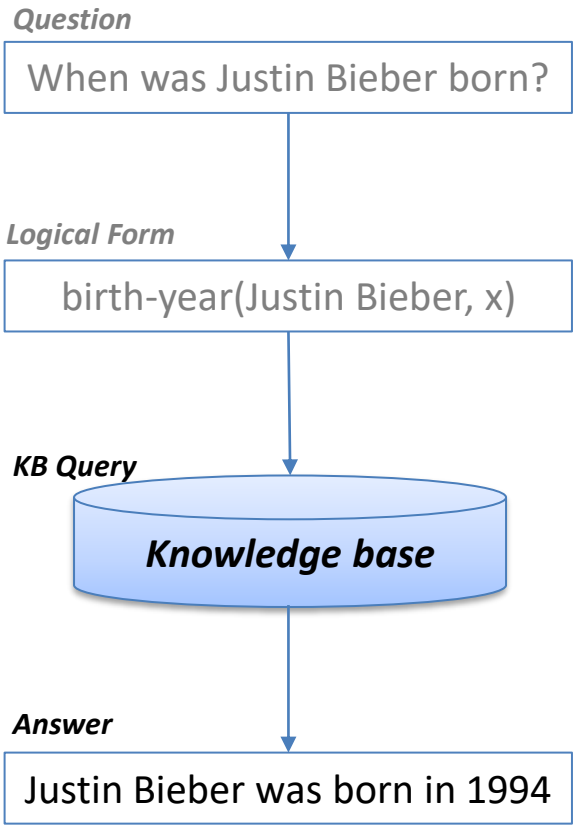
2. Decode with standard seq2seq model (training – teacher forcing)

1. Encode sentence with sequence models

2 Knowledge-based Question Answering

Semantic Parsing

Answering a natural language question by mapping it to a query over a structured database (formal representation of its meaning).



Answer questions that ask about one of the missing arguments in a triple

Subject	Predicate (relation)	Object
Justin Bieber	birth-year	1994
Frédéric Chopin	birth-year	1810
…	…	…

- DBPedia
- Freebase

How to produce the answer?

- *Seq2seq*
- *Template based generation*

Knowledge-based Question Answering

Pros and Cons of Knowledge-based QA

- *Logical Form instead of (direct) answer makes system robust*
- *Answer independent of question and parsing mechanism*

- *Constrained to queriable questions in Database Schema*
- *Difficult to find the well-structured training dataset*

Question Answering

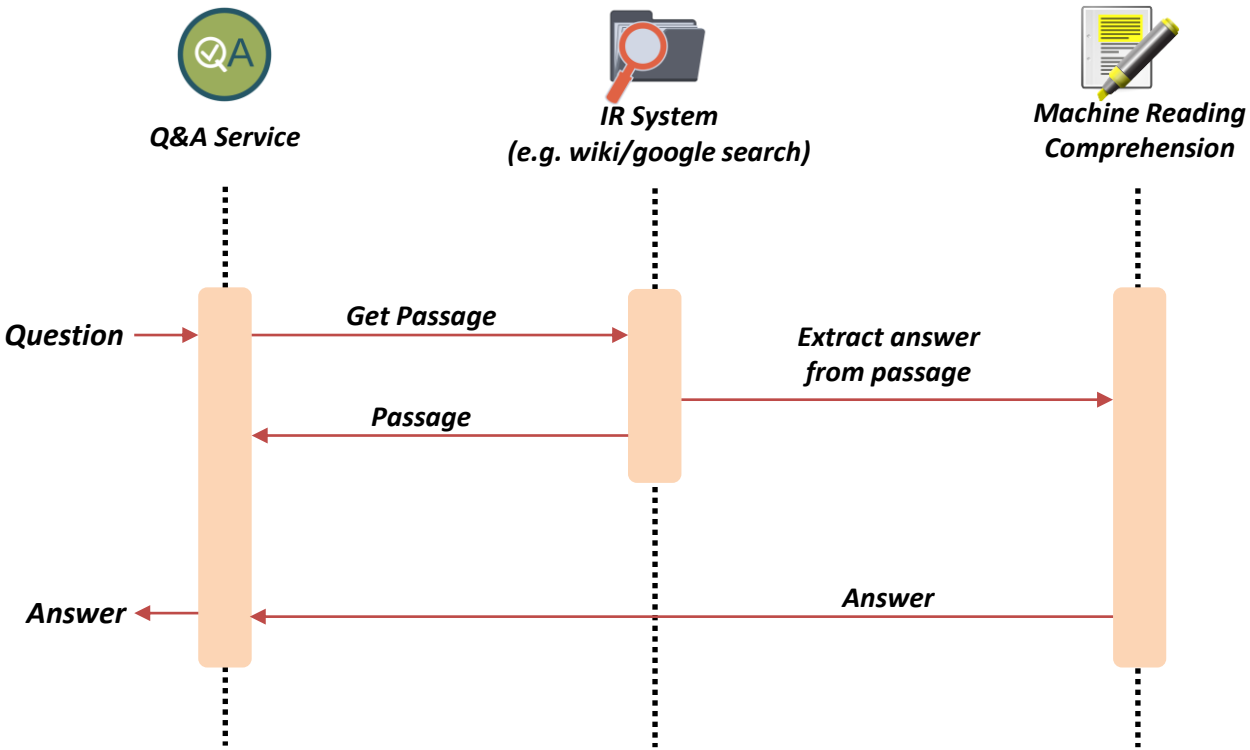
Research Areas in Question Answering

Research Area	Details
Knowledge-based QA (Semantic Parsing)	• Answer is a logical form, possible executed against a Knowledge Base • Context is a Knowledge Base
Information Retrieval-based QA • Answer sentence selection • Reading Comprehension	• Answer is a document, paragraph, sentence • Context is a corpus of documents or a specific document
Visual QA	• Answer is simple and factual • Context is one/multiple image(s)
Library Reference	• Answer is another question • Context is the structured knowledge available in the library and the librarians view of it.

Information Retrieval-based Question Answering

Information Retrieval-based Question Answering

Answering a user's question by finding short text segments, sentences, or documents on the web or collection of document



Reading Comprehension

Information Retrieval-based Question Answering

Answering a user's question by finding short text segments, sentences, or documents on the web or collection of document

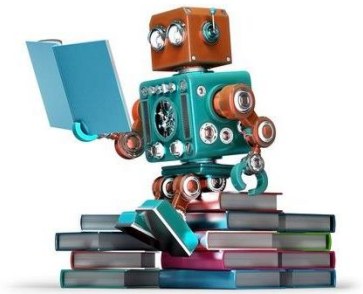
- ***Reading Comprehension and Answer Sentence Selection:***
  - *Finding an answer in a paragraph or a document*
  - *Picking a suitable sentence from a corpus that can be used to answer a question*

Reading Comprehension

Reading Comprehension

To answer these questions, you need to first gather information by collecting answer-related sentences from the article.

Can we teach this to machine?



Yes, we can!
Machine Comprehension of Text
(Burges 2013)

THE BOAT PARADE

The boats are floating along the lakeshore. It is the summer boat parade.

There are motor boats, rowboats and sailboats.

Jessica's favorite is the yellow motor boat with the flag. The rowboat decorated with flowers is Lisa's favorite. Tony likes the purple sailboat.

The boats float by one at a time. The people on the boats waive at the crowds. The crowds cheer the boats.

The boat parade is so much fun to watch. It is the best part of the summer.

Answer the Questions:

1. Where are the boats floating?

2. What kind of boats are there?

3. What is Lisa's favorite boat?

Reading Comprehension

Reading Comprehension

To answer these questions, you need to first gather information by collecting answer-related sentences from the article.

Can we teach this to machine?

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

THE BOAT PARADE

The boats are floating along the lakeshore. It is the summer boat parade.

There are motor boats, rowboats and sailboats.

Jessica's favorite is the yellow motor boat with the flag. The rowboat decorated with flowers is Lisa's favorite. Tony likes the purple sailboat.

The boats float by one at a time. The people on the boats waive at the crowds. The crowds cheer the boats.

The boat parade is so much fun to watch. It is the best part of the summer.

Answer the Questions:

1. Where are the boats floating?

2. What kind of boats are there?

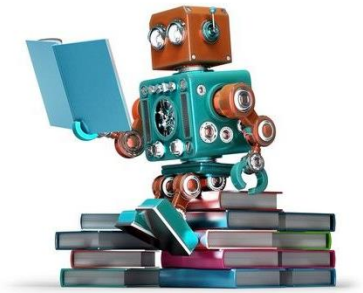
3. What is Lisa's favorite boat?

Reading Comprehension

Reading Comprehension

To answer these questions, you need to first gather information by collecting answer-related sentences from the article.

Why do we need to teach this?



The ability to comprehend text** will lead us to a **better search** and **solve lots of NLP problems!

THE BOAT PARADE

The boats are floating along the lakeshore. It is the summer boat parade.

There are motor boats, rowboats and sailboats.

Jessica's favorite is the yellow motor boat with the flag. The rowboat decorated with flowers is Lisa's favorite. Tony likes the purple sailboat.

The boats float by one at a time. The people on the boats waive at the crowds. The crowds cheer the boats.

The boat parade is so much fun to watch. It is the best part of the summer.

Answer the Questions:

1. Where are the boats floating?

2. What kind of boats are there?

3. What is Lisa's favorite boat?

Reading Comprehension

Corpora for Reading Comprehension

Dataset	Answer Type	Domain
MCTest (Richardson et al. 2013)	Multiple choice	Children's stories
CNN/Daily Mail (Hermann et al. 2015)	Spans	News
Children's book test (Hill et al. 2016)	Multiple choice	Children's stories
SQuAD (Rajpurkar et al., 2016)	Spans	Wikipedia
MS MARCO (Nguyen et al., 2016)	Free-from text, Unanswerable	Web Search
NewsQA (Trischler et al., 2017)	Spans	News
SearchQA (Dunn et al., 2017)	Spans	Jeopardy
TriviaQA (Joshi et al., 2017)	Spans	Trivia
RACE (Lai et al., 2017)	Multiple choice	Mid/High School Exams
Narrative QA (Kocisky et al., 2018)	Free-form text	Movie Scripts, Literature
SQuAD 2.0 (Rajpurkar et al., 2018)	Spans, Unanswerable	Wikipedia

Reading Comprehension

TriviaQA: A Large Scale Dataset for Reading Comprehension

TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

The full dataset is coming soon. Here's a sneak peek! The evidence documents come from two domains -- Wikipedia and the web. Click on the "Evidence" button to see the document for each question.

QuestionId	Question	Answer	Web	Wikipedia
qw_3199	Miami Beach in Florida borders which ocean?	Atlantic	Evidence	Evidence
bt_1255	What was the occupation of Lovely Rita according to the song by the Beatles	Traffic Warden	Evidence	Evidence
qg_77	Who was Poopdeck Pappys most famous son?	Popeye	Evidence	Evidence
wh_1026	The Nazi regime was Germany's Third Reich; which was the first Reich?	HOLY ROMAN EMPIRE	Evidence	Evidence
bb_1342	At which English racecourse did two horses collapse and die in the parade ring due to electrocution, in February 2011?	Newbury	Evidence	Evidence
wh_2759	Which type of hat takes its name from an 1894 novel by George Du Maurier where the title character has the surname O'Ferrall ?	TRILBY	Evidence	Evidence
sfq_8522	What was the Elephant Man's real name?	Joseph Merrick	Evidence	Evidence

http://nlp.cs.washington.edu/triviaqa/sample.html

Reading Comprehension

SQuAD: Stanford Question Answering Dataset

Victoria_(Australia)

The Stanford Question Answering Dataset

The economy of Victoria is highly diversified: service sectors including financial and property services, health, education, wholesale, retail, hospitality and manufacturing constitute the majority of employment. Victoria's total gross state product (GSP) is ranked second in Australia, although Victoria is ranked fourth in terms of GSP per capita because of its limited mining activity. Culturally, Melbourne is home to a number of museums, art galleries and theatres and is also described as the "sporting capital of Australia". The Melbourne Cricket Ground is the largest stadium in Australia, and the host of the 1956 Summer Olympics and the 2006 Commonwealth Games. The ground is also considered the "spiritual home" of Australian cricket and Australian rules football, and hosts the grand final of the Australian Football League (AFL) each year, usually drawing crowds of over 95,000 people. Victoria includes eight public universities, with the oldest, the University of Melbourne, having been founded in 1853.

What kind of economy does Victoria have?
Ground Truth Answers: diversified | highly diversified | highly diversified
Prediction: highly diversified

Where according to gross state product does Victoria rank in Australia?
Ground Truth Answers: second | second | second
Prediction: second

At what rank does GPS per capita set Victoria?
Ground Truth Answers: fourth | fourth | fourth
Prediction: fourth

What city in Victoria is called the sporting capital of Australia?
Ground Truth Answers:
Melbourne | Melbourne | Melbourne
Prediction: Melbourne

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

Reading Comprehension

A Generic Neural Model for Reading Comprehension

***Step1**: For both documents and questions, convert words to word vectors*

Document (D)

A partly submerged glacier cave on Perito Moreno Glacier. The ice facade is approximately 60 m high. Ice formations in the Titlis glacier cave. A glacier cave is a cave formed within the ice of a glacier. Glacier caves are often called ice caves, but the latter term is properly used to describe bedrock caves that contain year-round ice

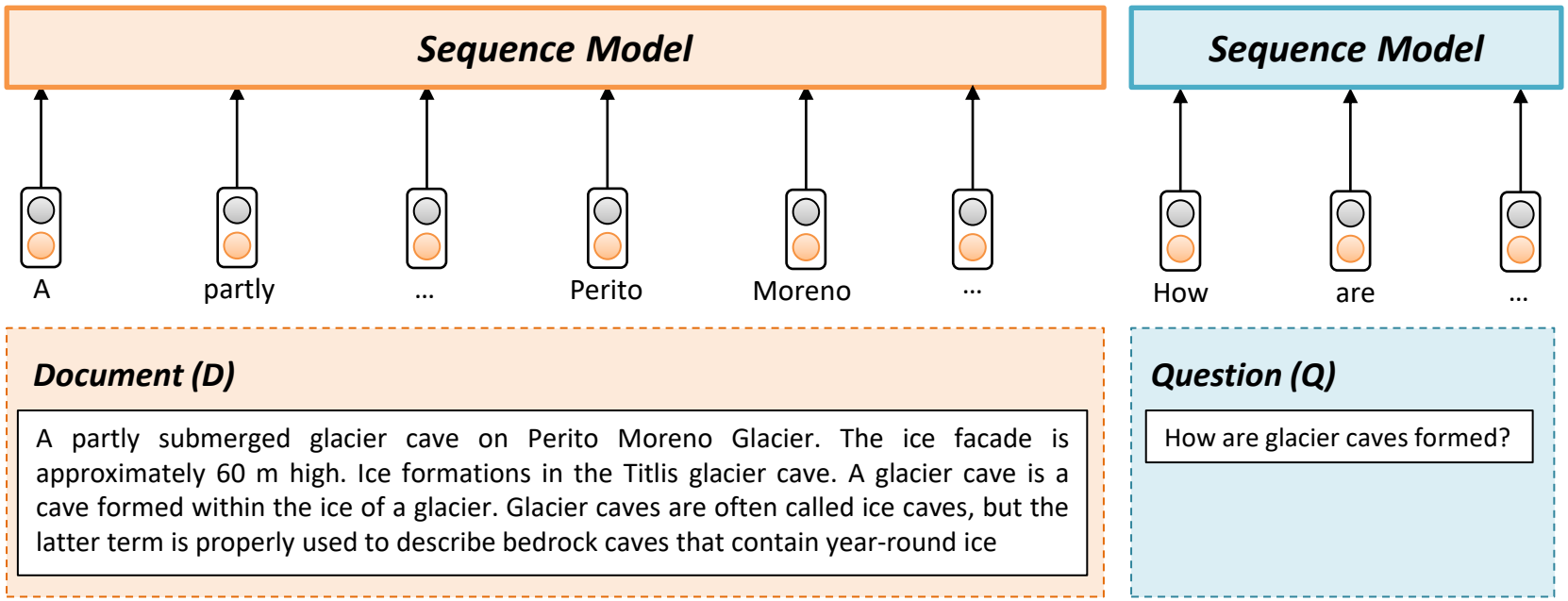
Question (Q)

How are glacier caves formed?

Reading Comprehension

A Generic Neural Model for Reading Comprehension

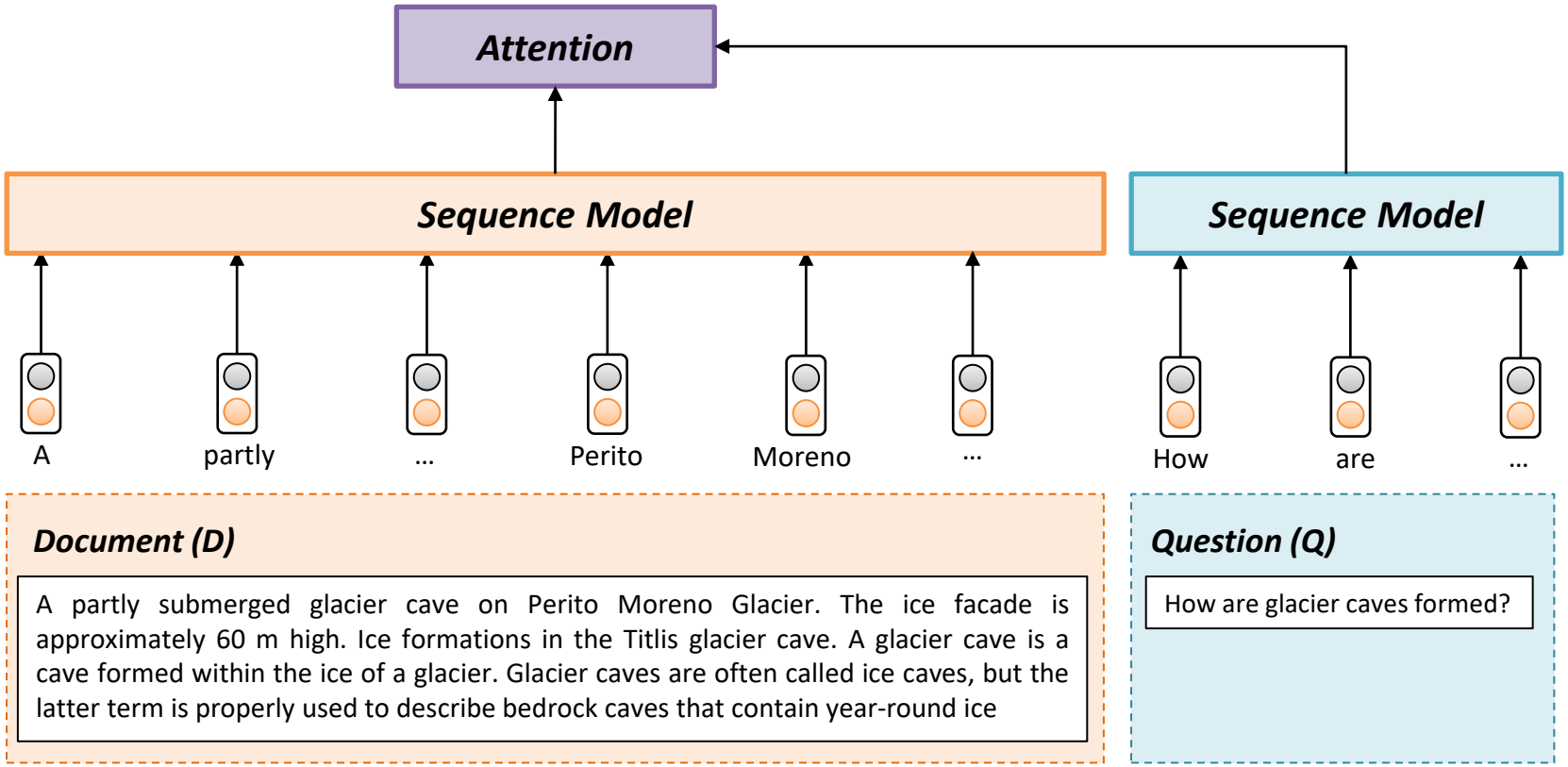
Step2: *Encode context (documents) and question with sequence models*



Reading Comprehension

A Generic Neural Model for Reading Comprehension

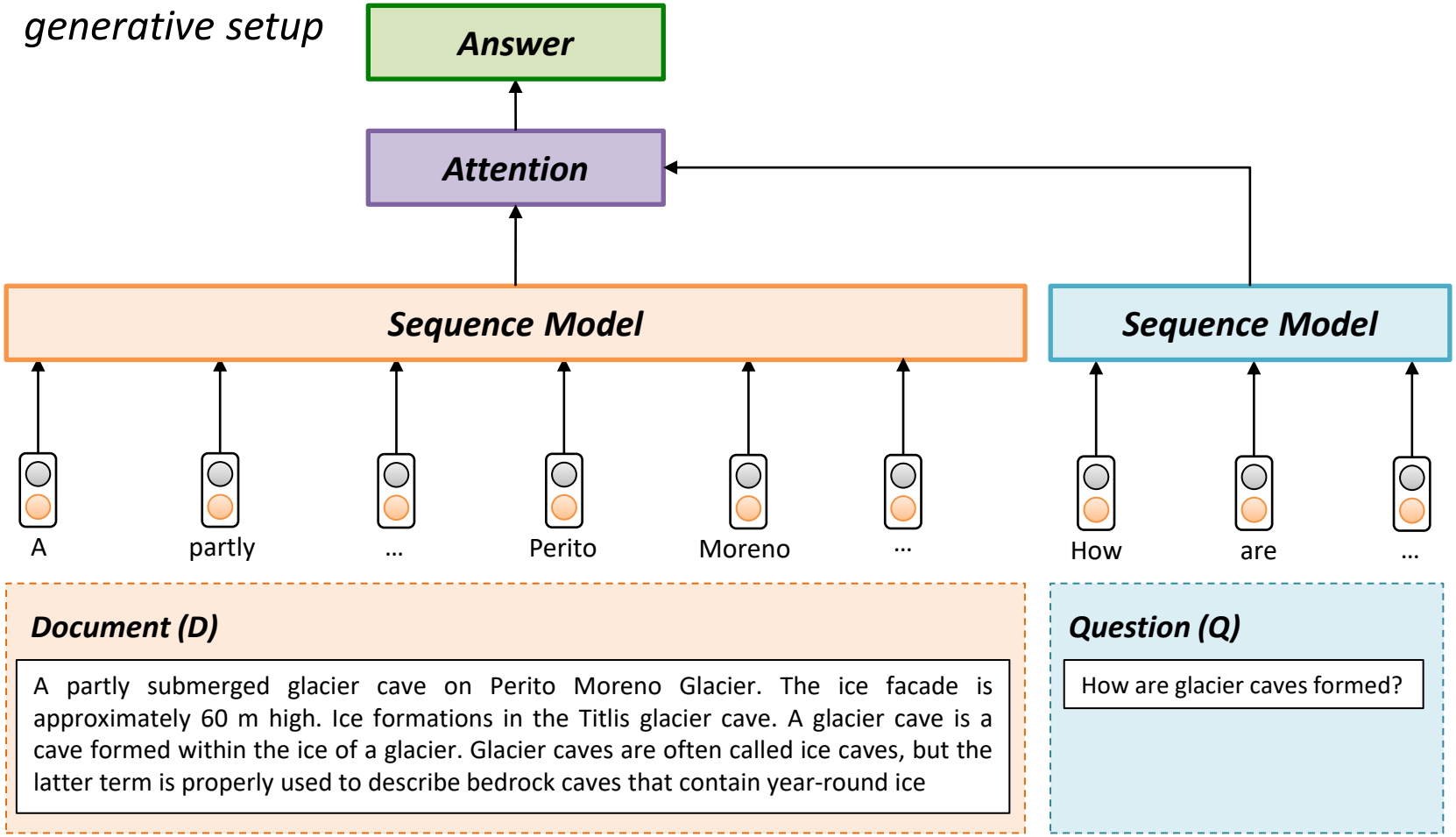
***Step3**: Combine context (documents) and question with an attention*



Reading Comprehension

A Generic Neural Model for Reading Comprehension

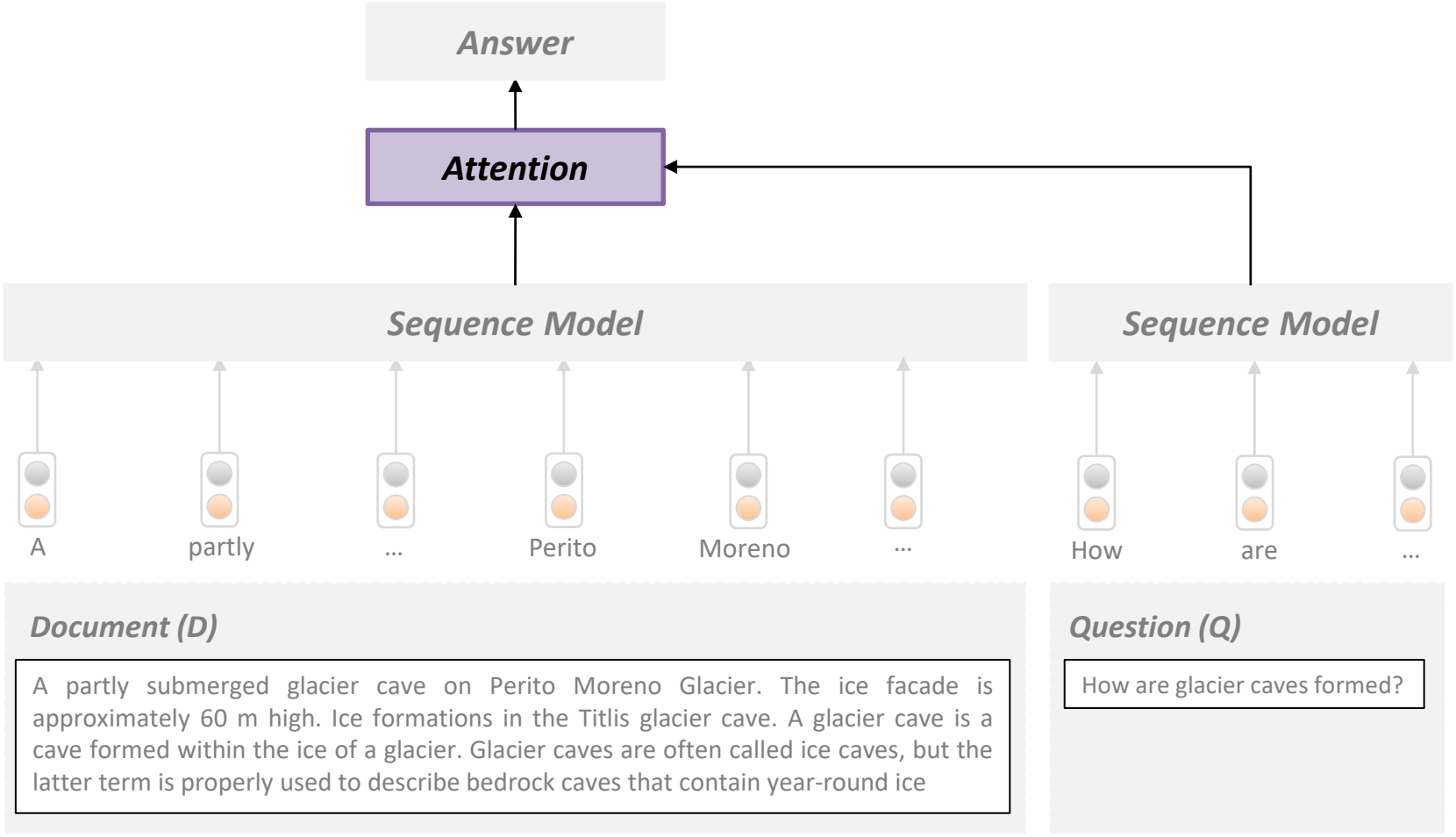
***Step4**: Select answer from attention map by using a classifier or with generative setup*



Reading Comprehension

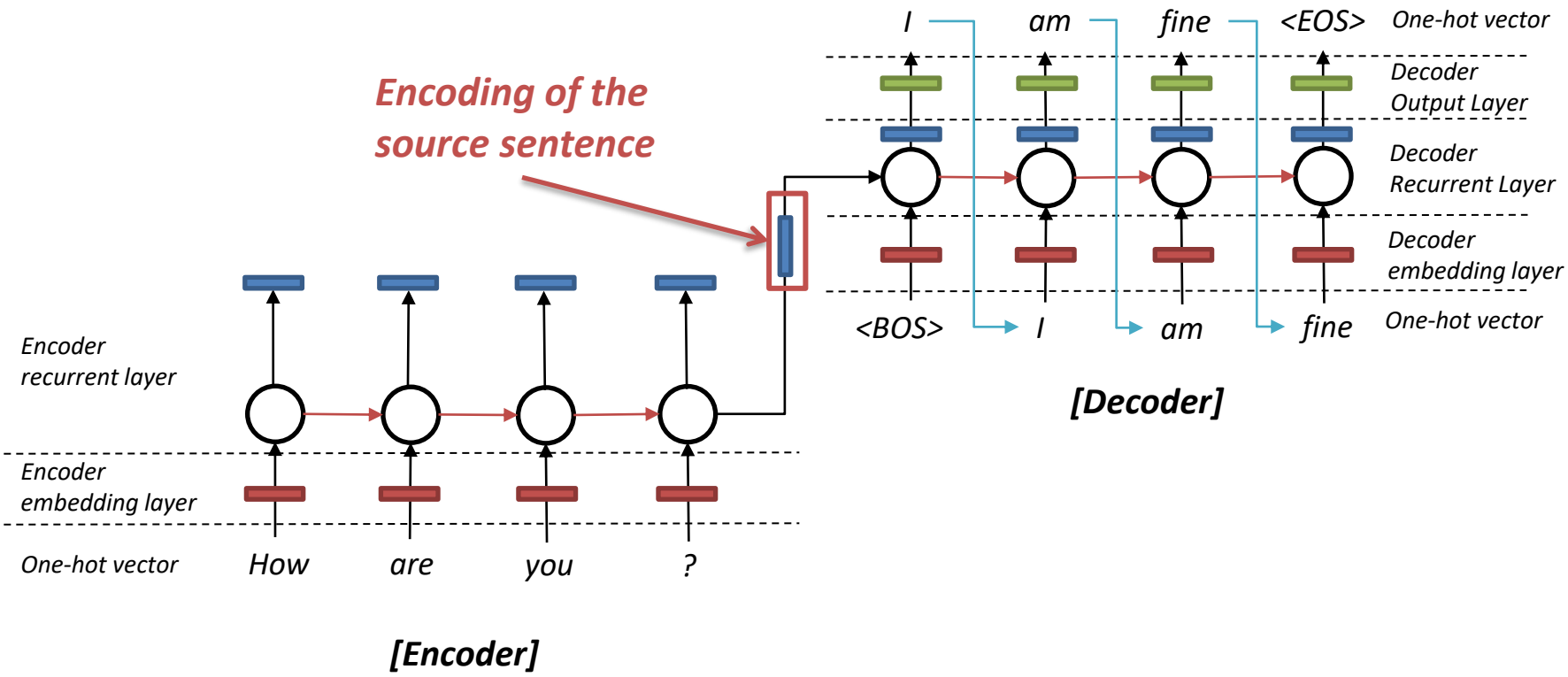
A Generic Neural Model for Reading Comprehension

*What is the **Attention**? Why we need this?*



Attention

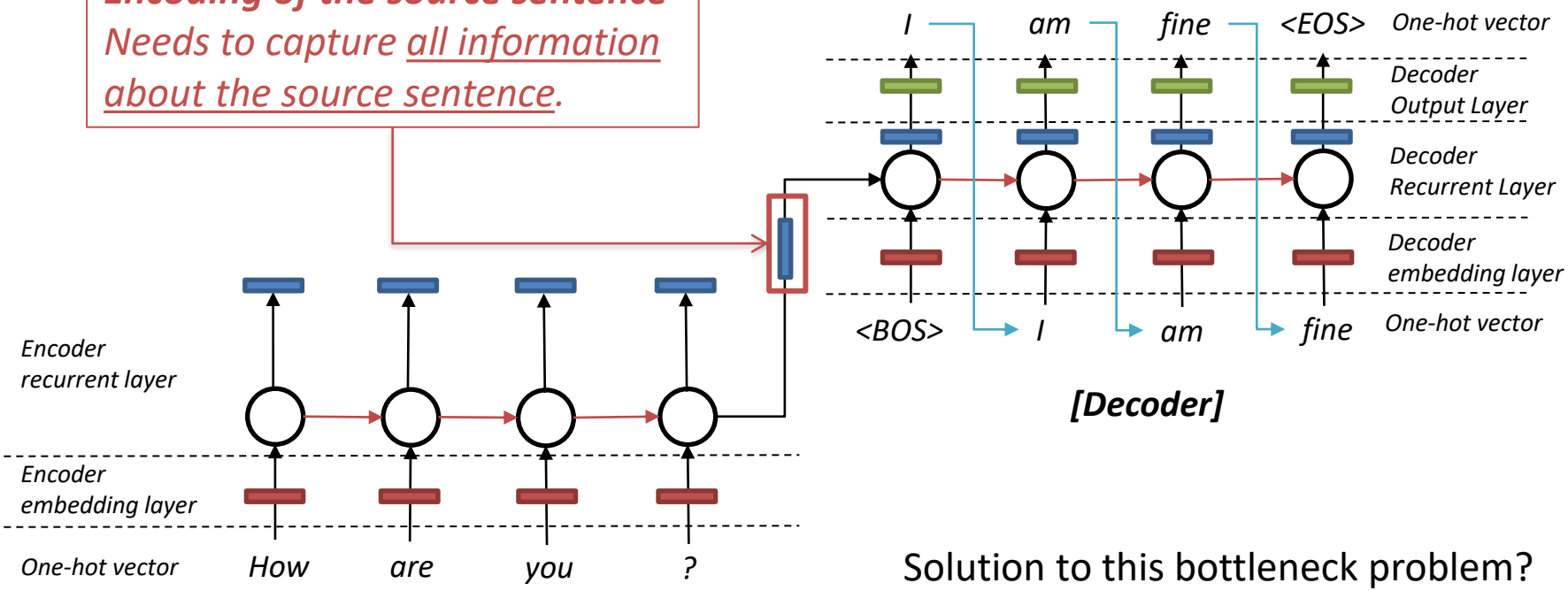
Seq2Seq Model: Recap



Attention

Seq2Seq Model: the bottleneck problem

Encoding of the source sentence
Needs to capture all information about the source sentence.



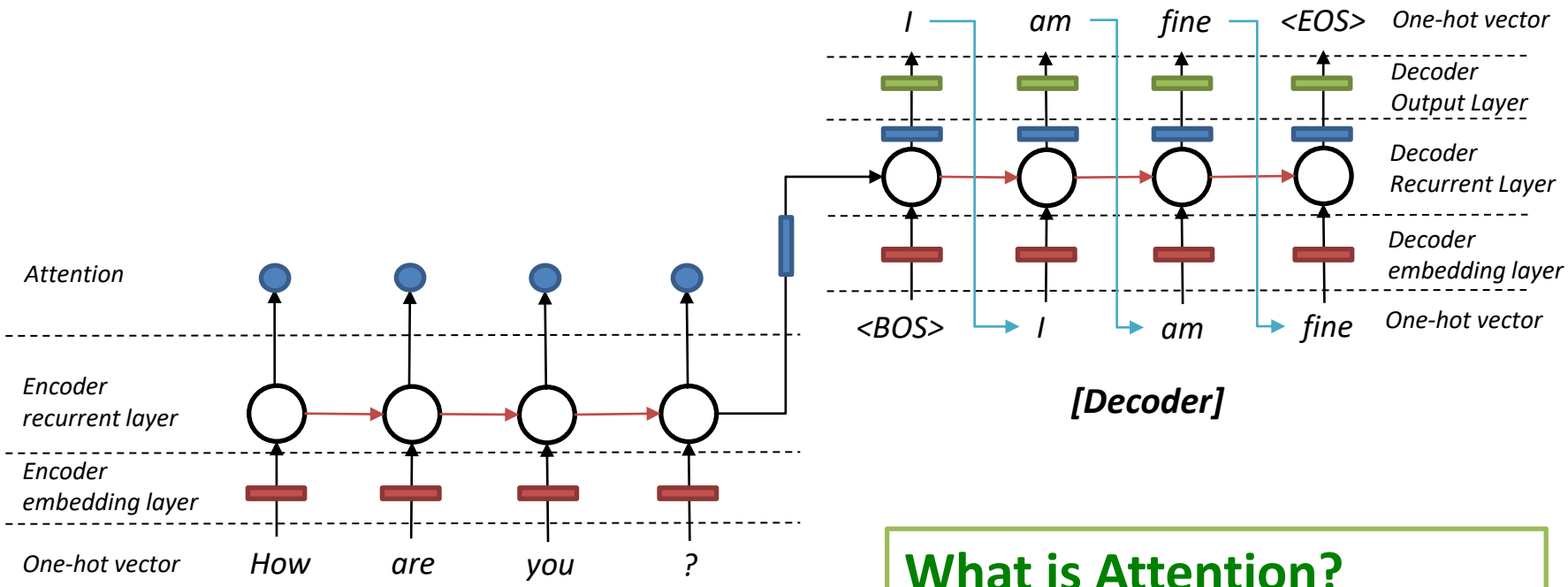
+RNN drawback!
Vanishing Gradient

Solution to this bottleneck problem?

Attention!

Attention

Seq2Seq with Attention

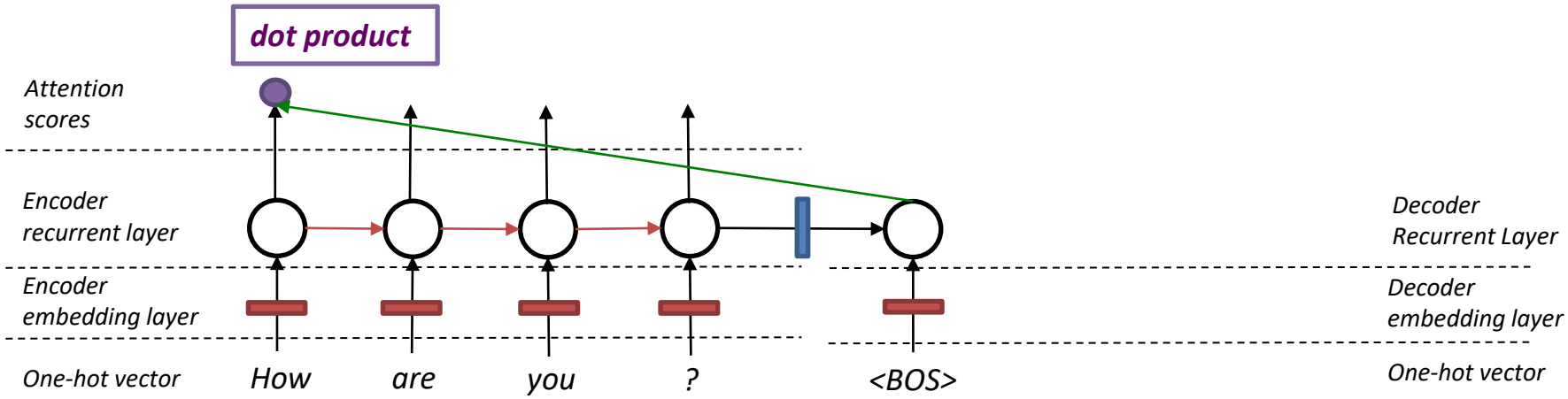


What is Attention?

*On **each step of the decoder**, use **direct connection to the encoder** to **focus on a particular part** of the input sequence*

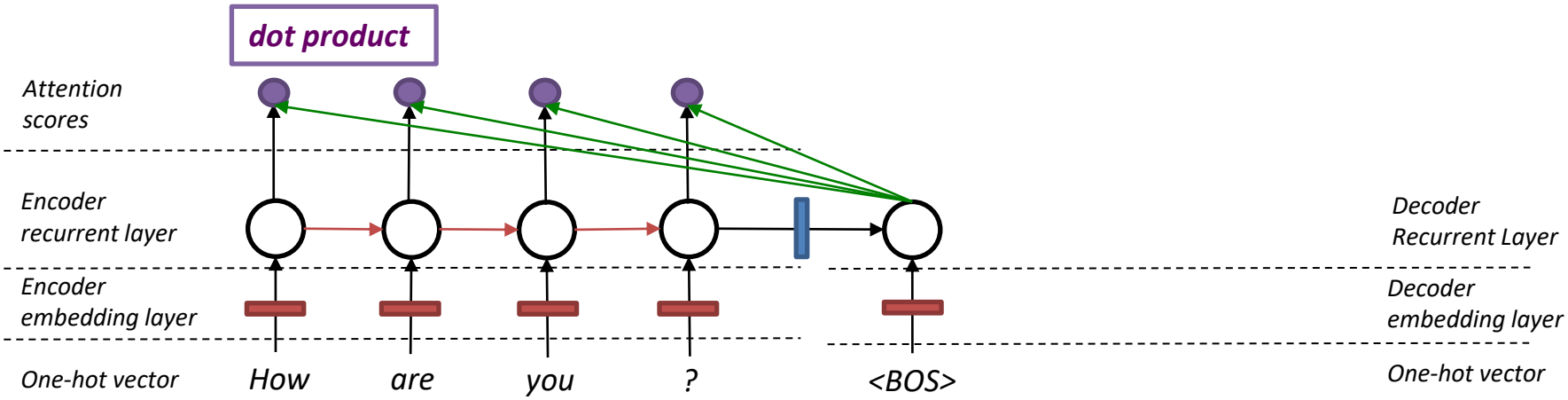
Attention

Seq2Seq with Attention



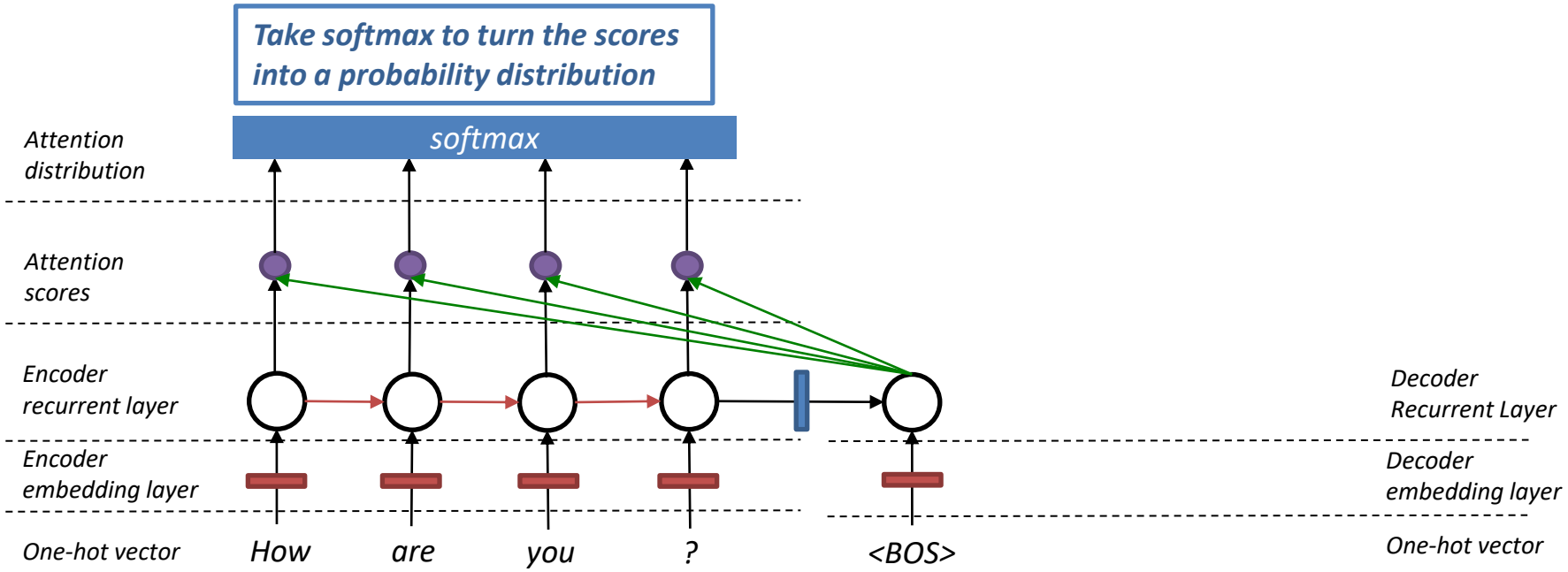
Attention

Seq2Seq with Attention



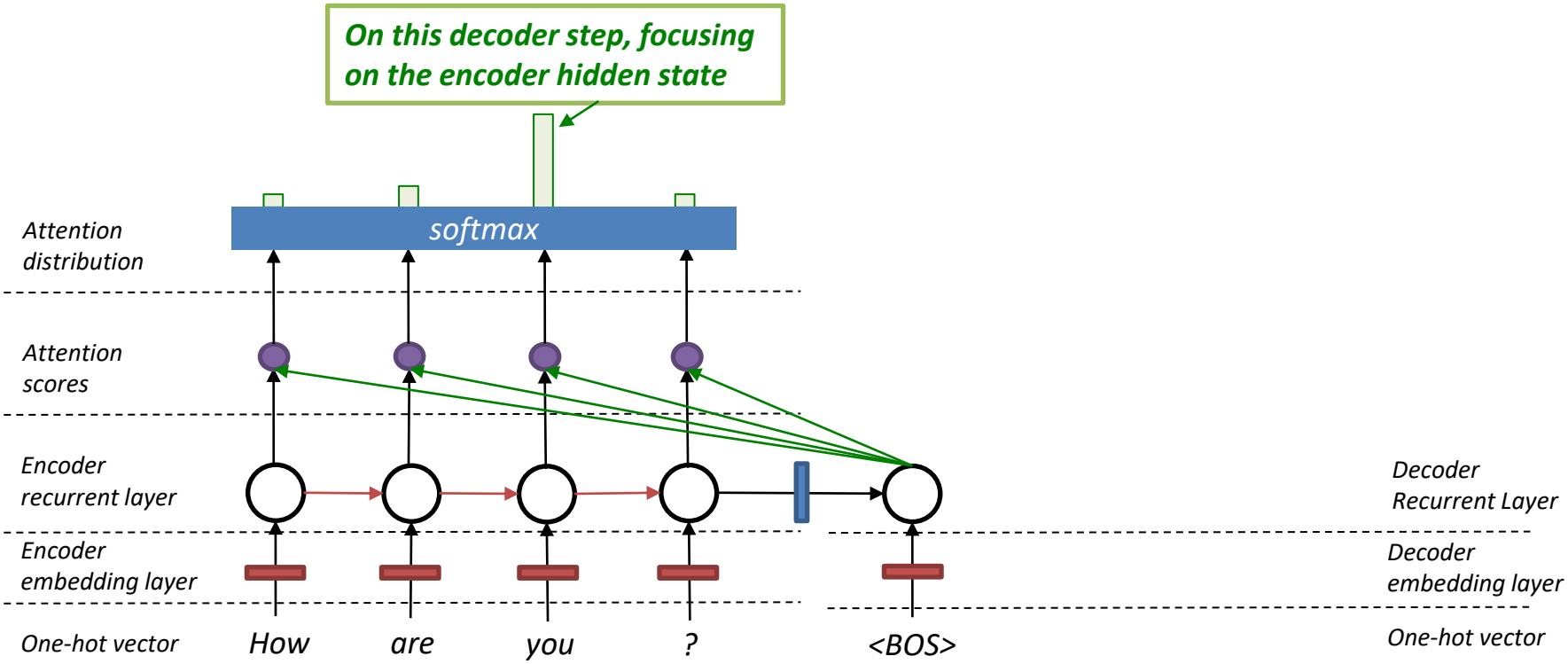
Attention

Seq2Seq with Attention



Attention

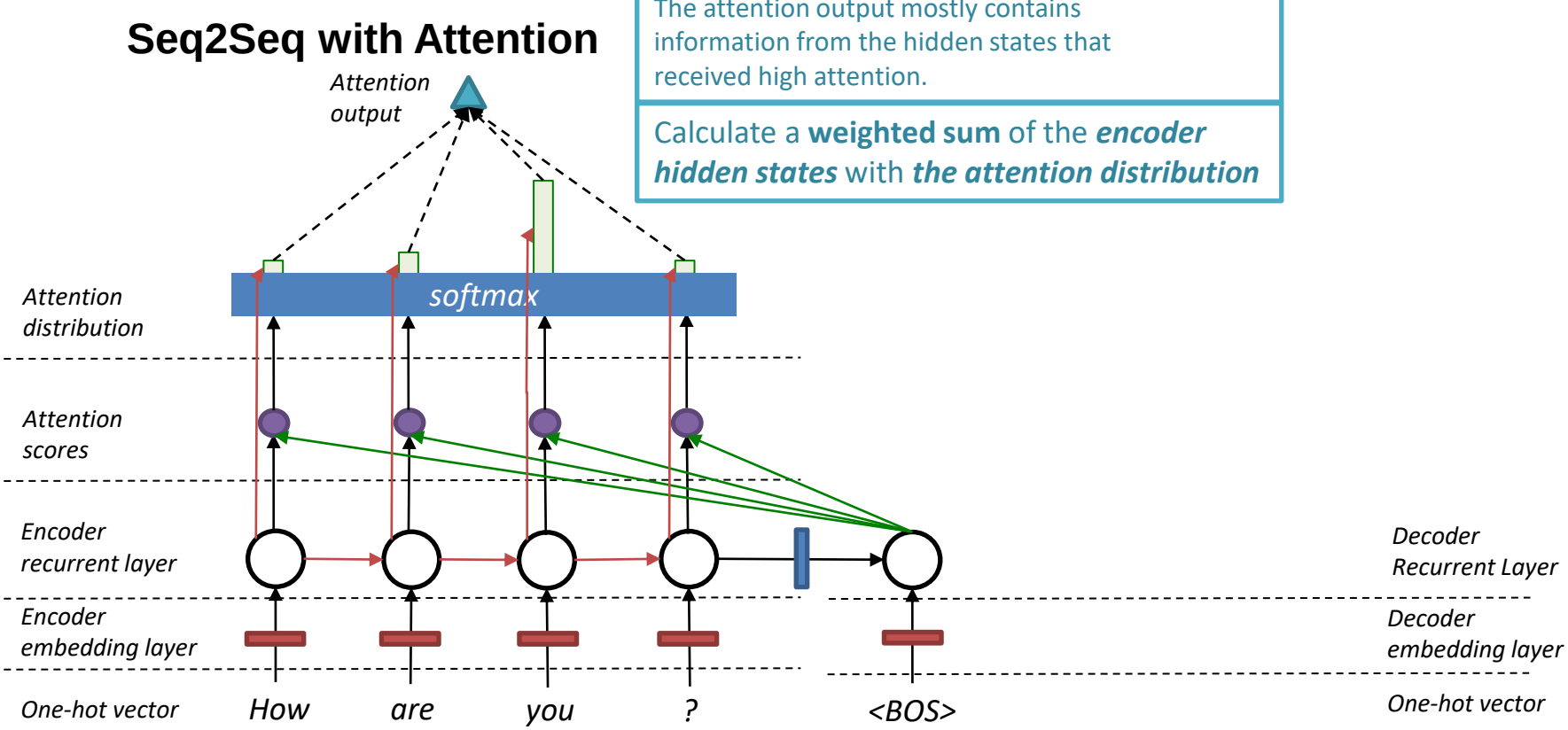
Seq2Seq with Attention



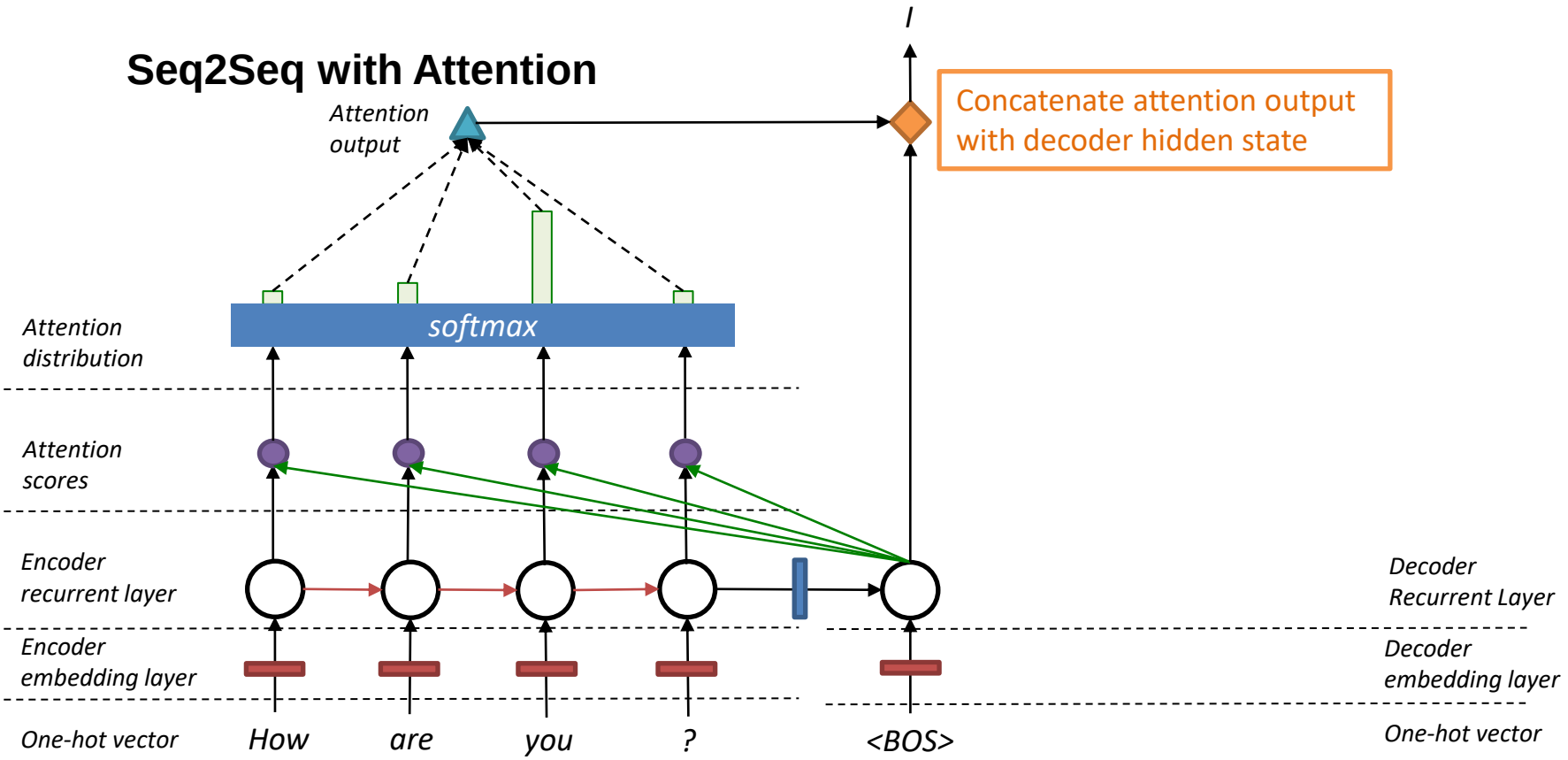
Attention

The attention output mostly contains information from the hidden states that received high attention.

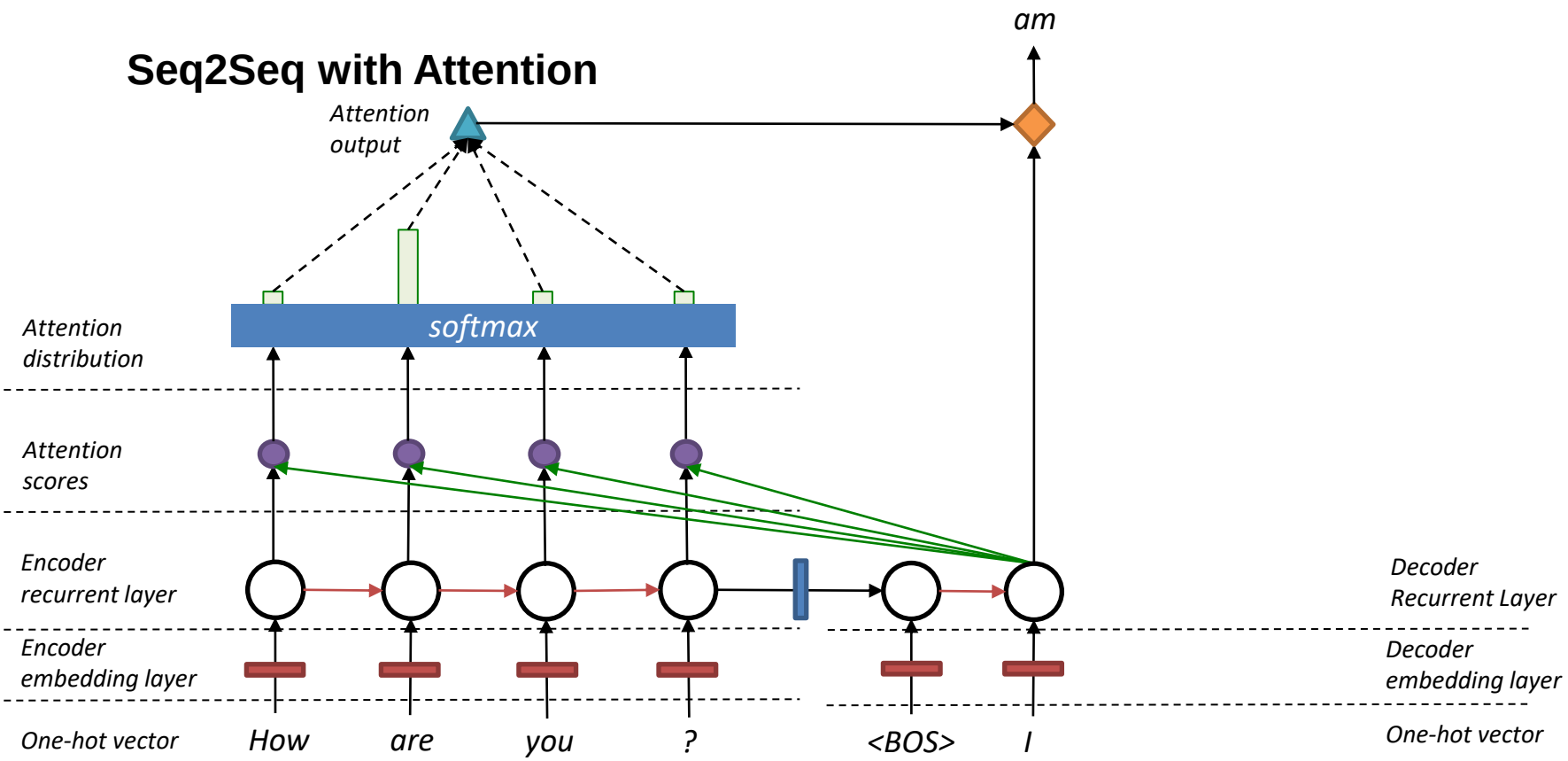
Calculate a **weighted sum** of the ***encoder hidden states*** with ***the attention distribution***



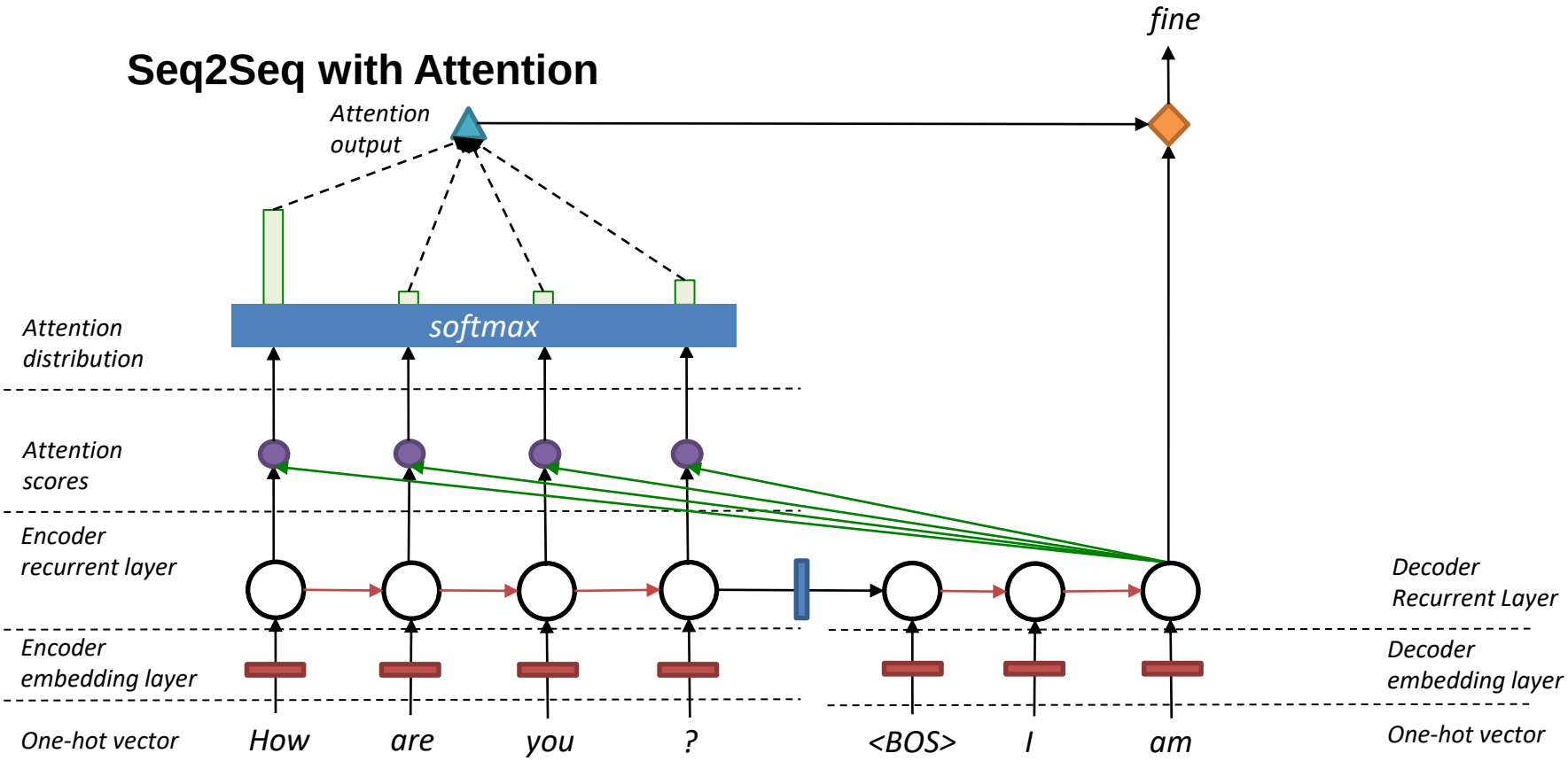
Attention


Seq2Seq with Attention
Attention output
I
Concatenate attention output with decoder hidden state
softmax
Attention distribution
Attention scores
Encoder recurrent layer
Encoder embedding layer
One-hot vector
How
are
you
?
<BOS>
Decoder Recurrent Layer
Decoder embedding layer
One-hot vector

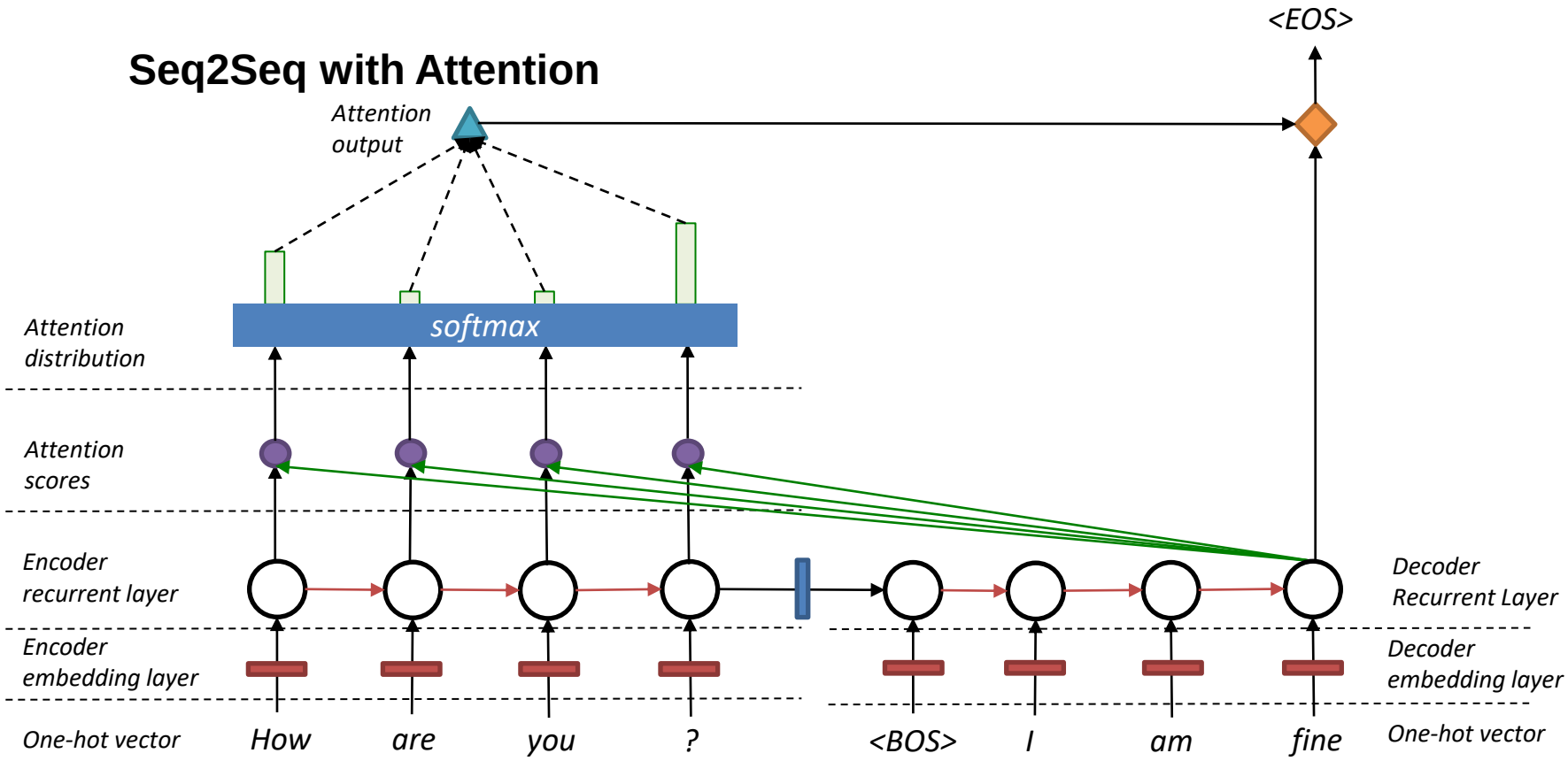
Attention


Seq2Seq with Attention
am
Attention output
softmax
Attention distribution
Attention scores
Encoder recurrent layer
Encoder embedding layer
One-hot vector
How
are
you
?
<BOS>
I
Decoder Recurrent Layer
Decoder embedding layer
One-hot vector

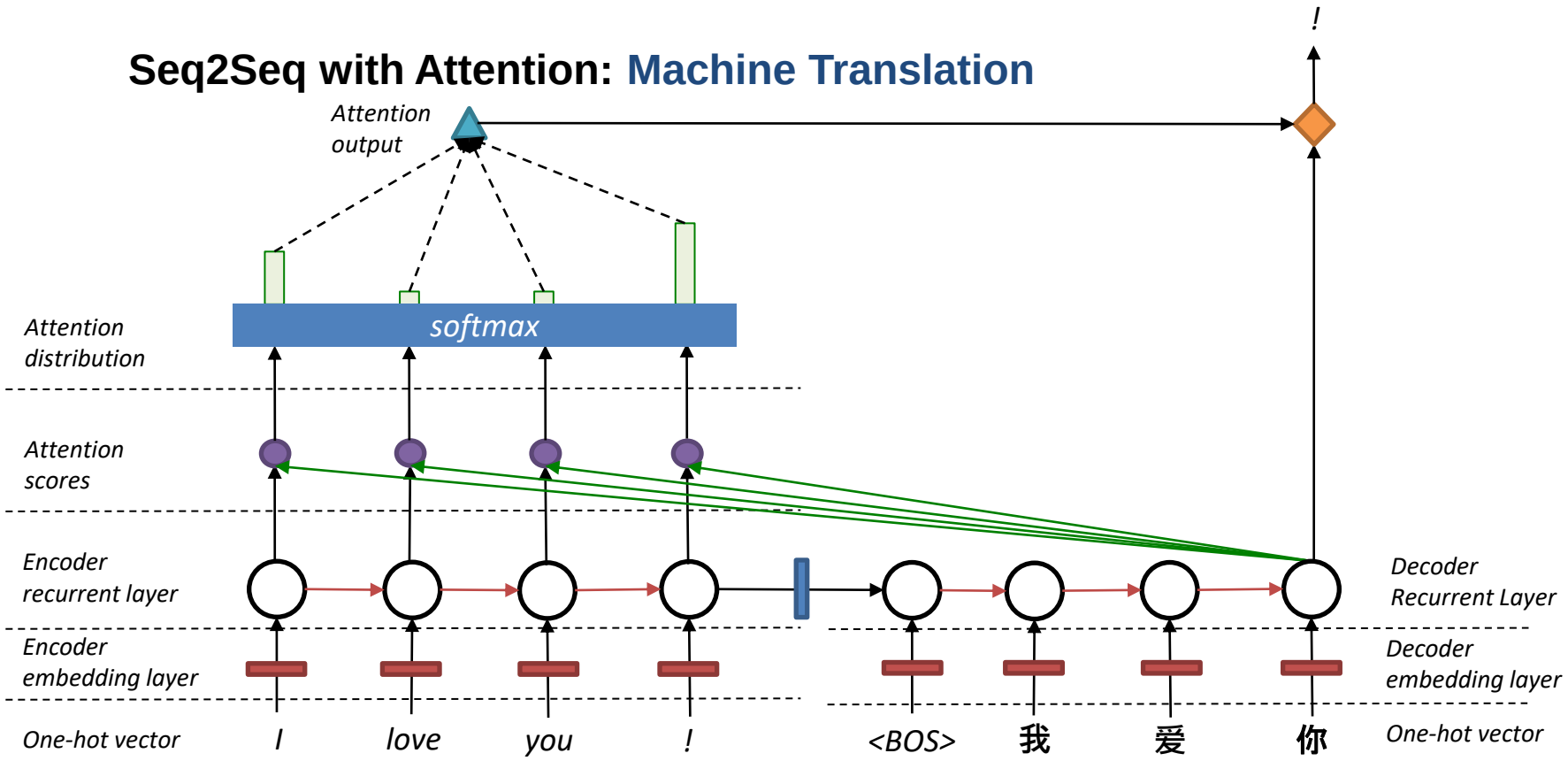
Attention


Seq2Seq with Attention
fine
Attention output
softmax
Attention distribution
Attention scores
Encoder recurrent layer
Encoder embedding layer
One-hot vector
How
are
you
?
<BOS>
I
am
Decoder Recurrent Layer
Decoder embedding layer
One-hot vector

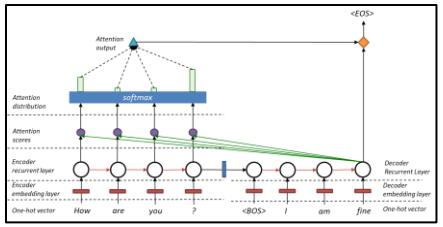
Attention


Seq2Seq with Attention
<EOS>
Attention output
softmax
Attention distribution
Attention scores
Encoder recurrent layer
Encoder embedding layer
One-hot vector
How
are
you
?
<BOS>
I
am
fine
Decoder Recurrent Layer
Decoder embedding layer
One-hot vector

Attention


Seq2Seq with Attention: Machine Translation
Attention output
!
softmax
Attention distribution
Attention scores
Encoder recurrent layer
Decoder Recurrent Layer
Encoder embedding layer
Decoder embedding layer
One-hot vector
I
love
you
!
<BOS>
我
爱
你
One-hot vector

Attention



Seq2Seq with Attention (Equations)

- *Encoder hidden states:* $h_1, \ldots, h_N \in \mathbb{R}^h$
- *Decoder hidden state:* $s_t \in \mathbb{R}^h$ *(on timestep t)*

1. *Attention score* $\boldsymbol{e}^t$: $\boldsymbol{e}^t = [\boldsymbol{s}_t^T \boldsymbol{h}_1, \ldots, \boldsymbol{s}_t^T \boldsymbol{h}_N] \in \mathbb{R}^N$ *(for timestep t)*

2. *Use softmax to get the attention distribution,* α^t *(for timestep t)*
   (this is a probability distribution and sums to 1)

$$\alpha^t = \text{softmax}(\boldsymbol{e}^t) \in \mathbb{R}^N$$

3. *Attention Output: Use* α^t *to take a weighted sum of the encoder hidden states*

$$\boldsymbol{a}_t = \sum_{i=1}^{N} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

4. *Then, concatenate the attention output* $\boldsymbol{a}_t$ *with the decoder hidden state* s_t *and proceed as in the non-attention seq2seq model*

$$[\boldsymbol{a}_t; \boldsymbol{s}_t] \in \mathbb{R}^{2h}$$

Attention

Why we use Attention? The benefit!

Improve performance

- *Allow decoder to focus on certain parts of the source*

Solving the bottleneck problem

- *Allow decoder to directly look at the source (input)*

Reducing vanishing gradient problem

- *Provide shortcut to faraway states*

Providing some interpretability

- *Inspect attention distribution, and show what the decoder was focusing on*

Attention

Attention is now a general component in Deep Learning NLP

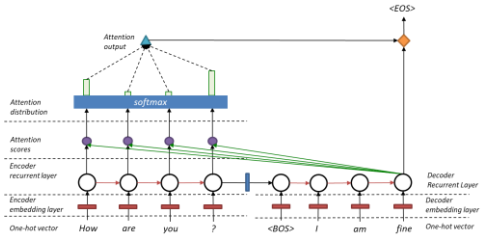
Attention is great way to improve the sequence to sequence model.

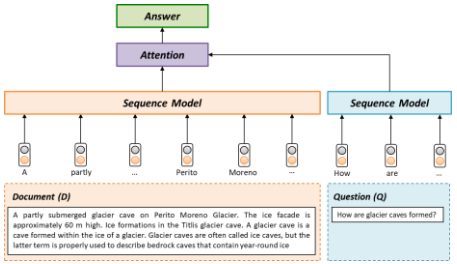
You can use attention in many architecture (not just seq2seq) and many NLP tasks (not just dialog system/NLG, Translation)

More general definition of attention:

Given a set of vector values, and a vector query, attention is a technique to compute a weighted sum of the values, dependent on the query.

*For example, in the seq2seq + attention model, each decoder hidden state (**query**) attends to all the encoder hidden states (**values**).*





Attention

Attention variants

There are several ways to compute attention score.

- *Encoder hidden states:* $h_1, \ldots, h_N \in \mathbb{R}^h$
- *Decoder hidden state:* $s_t \in \mathbb{R}^h$ *(on timestep t)*

Attention Name	Attention score function	Reference
Content-base	$score(s_t, h_i) = cosine[s_t, h_i]$	Graves 2014
Dot-product	$score(s_t, h_i) = s_t^\top h_i$	Luong 2015
Scaled Dot-product	$score(s_t, h_i) = \frac{s_t^\top h_i}{\sqrt{n}}$ *NOTE: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani 2017
Additive	$score(s_t, h_i) = v_a^\top \tanh(W_a[s_t; h_i])$	Vaswani 2017
General	$score(s_t, h_i) = s_t^\top \mathrm{W}_a h_i$ *NOTE: where W_a is a trainable weight matrix in the attention layer.	Luong 2015
Location-based	$a_{t,\mathrm{i}} = softmax(W_a s_t)$ *Note: This simplifies the softmax alignment to only depend on the target position.	Luong 2015

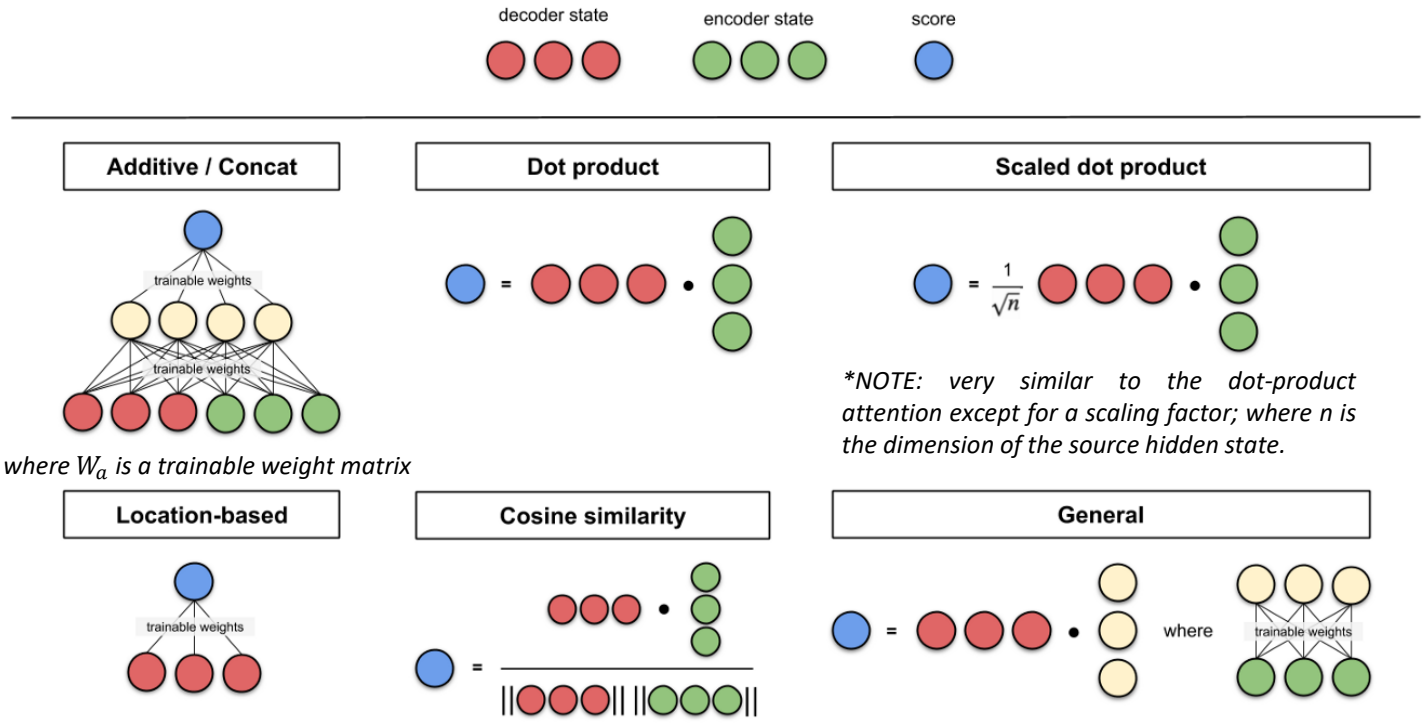
**The papers (Luong 2015 and Vaswani 2017) can be found in the canvas content page*

Attention

Attention variants

There are several ways to compute attention score.

- *Encoder hidden states:* $h_1, \ldots, h_N \in \mathbb{R}^h$
- *Decoder hidden state:* $s_t \in \mathbb{R}^h$ *(on timestep t)*



**Note: This simplifies the softmax alignment to only depend on the target position.*

Attention

Categories of Attention Mechanism

A summary of broader categories of attention mechanisms

Name	*Definition*	*Citation*
Global or Local	• Global: Attending to the entire input state space. • Local: Attending to the part of input state space (i.e. a patch of the input image.)	**Luong 2015**
Self-Attention	Relating different positions of the same input sequence. Theoretically the self-attention can adopt any attention score functions, but just replace the target sequence with the same input sequence.	**Cheng 2016**

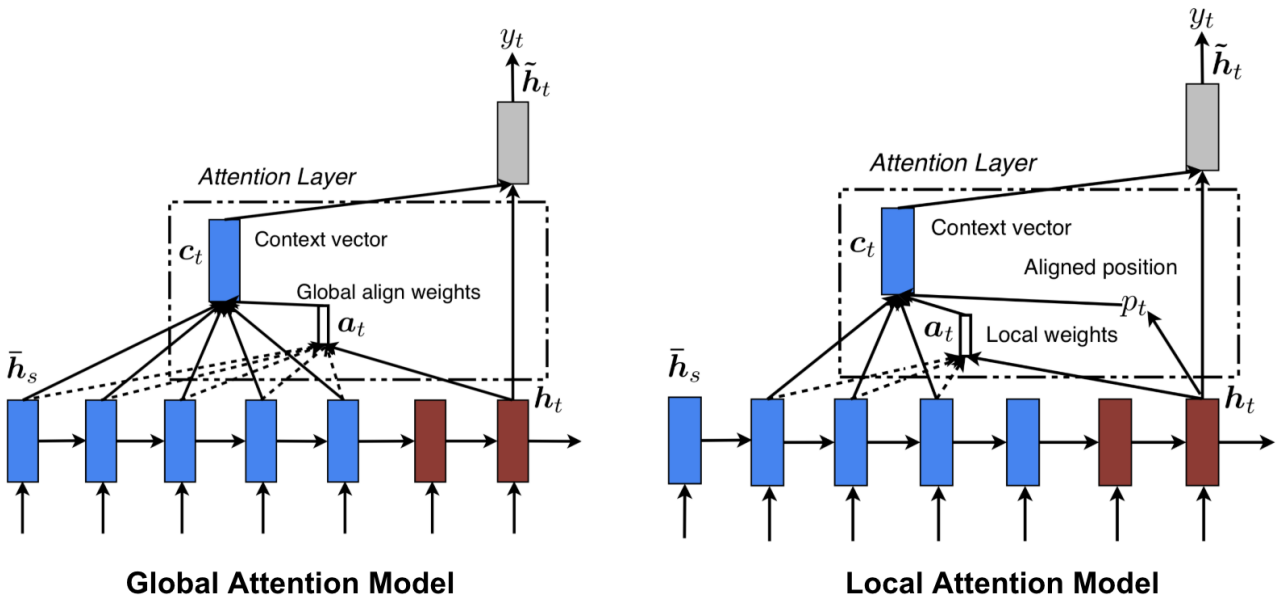
**The papers (Luong 2015 and Cheng 2016) can be found in the canvas content page*

Attention

Categories of Attention Mechanism (1)

Global/Local Attention

- *Global: Attending to the entire input state space.*
- *Local: Attending to the part of input state space*

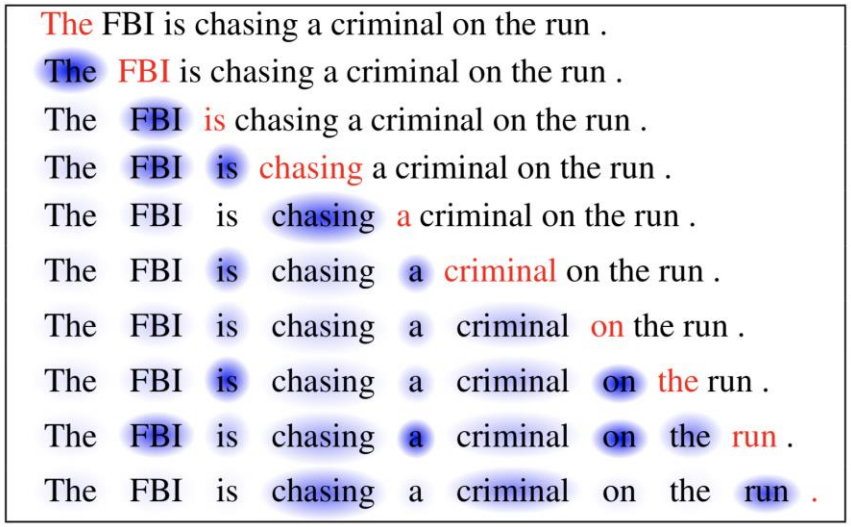


Attention

Categories of Attention Mechanism (2)

Self-Attention

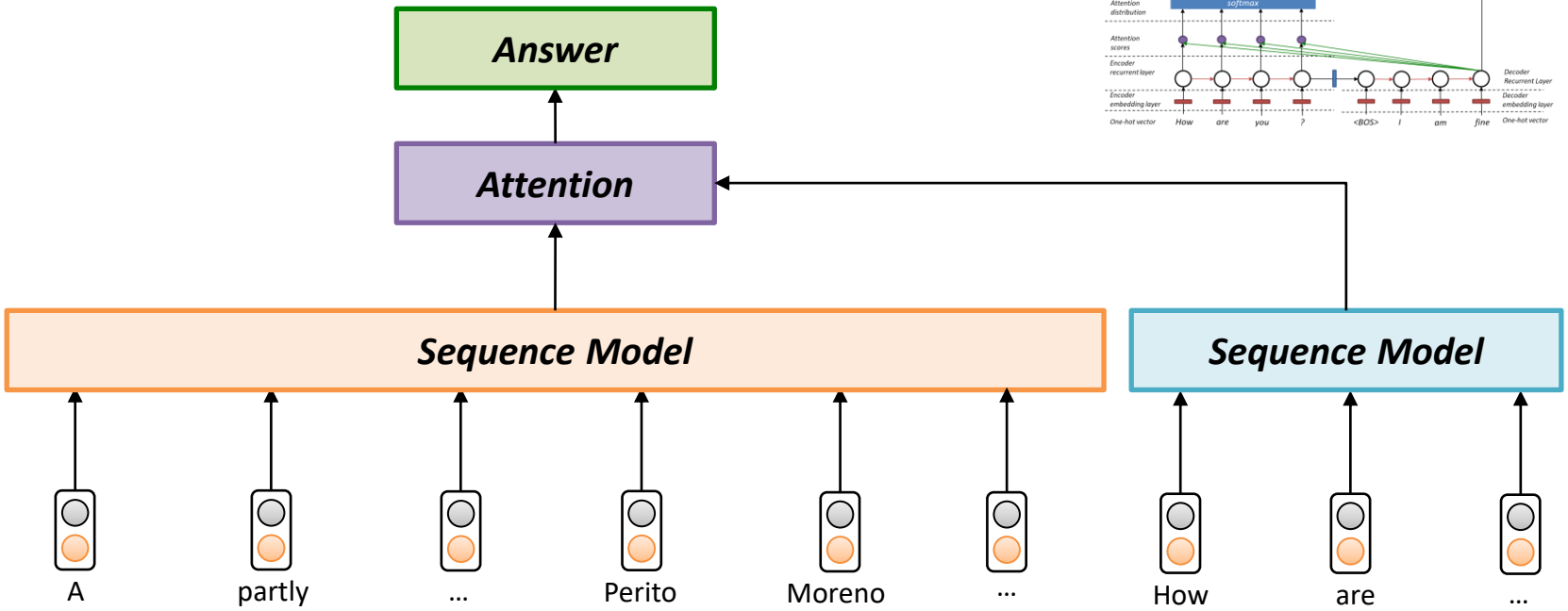
The long short-term memory network (Cheng et al., 2016) paper used self-attention to do machine reading. In the example below, the self-attention mechanism enables us to learn the correlation between the current words and the previous part of the sentence.



The current word is in red and the size of the blue shade indicates the activation level.

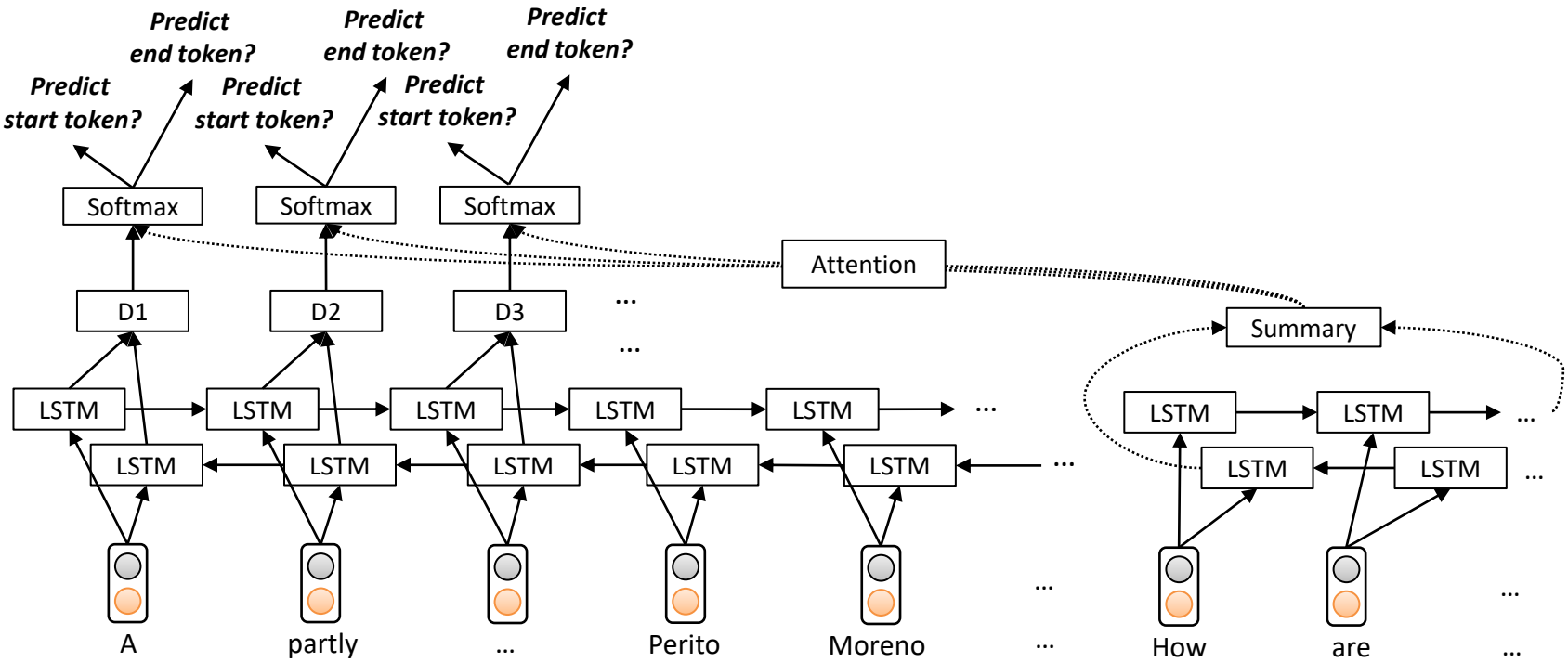
Reading Comprehension with Attention

A Generic Neural Model for Reading Comprehension



Reading Comprehension with Attention

Bi-LSTM for Reading Comprehension



Document (D)

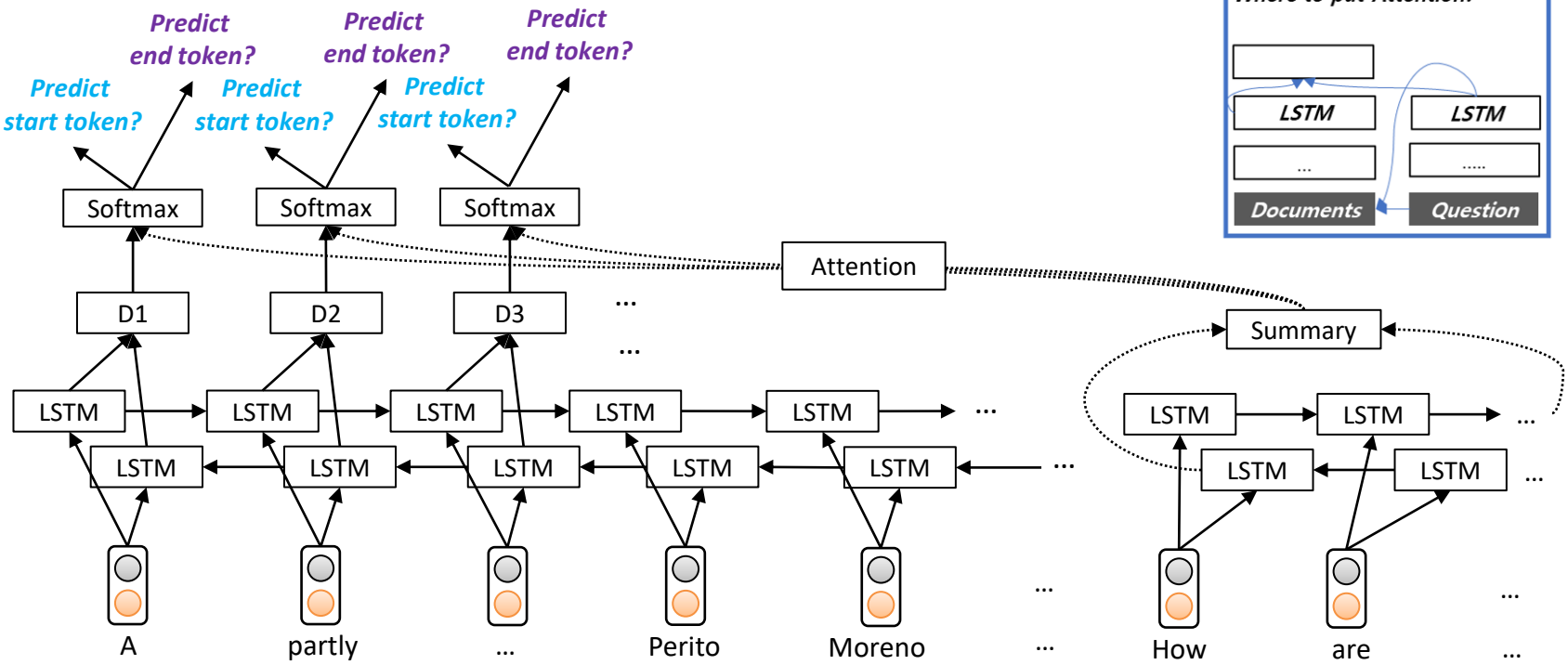
A partly submerged glacier cave on Perito Moreno Glacier. The ice facade is approximately 60 m high. Ice formations in the Titlis glacier cave. A glacier cave is a cave formed within the ice of a glacier. Glacier caves are often called ice caves, but the latter term is properly used to describe bedrock caves that contain year-round ice

Question (Q)

How are glacier caves formed?

Reading Comprehension with Attention

Bi-LSTM for Reading Comprehension with Attention



Document (D)

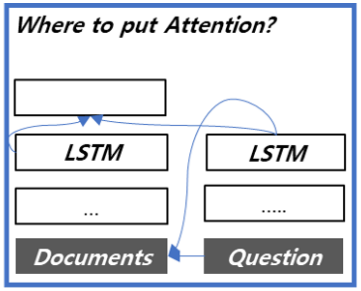
A partly submerged glacier cave on Perito Moreno Glacier. The ice facade is approximately 60 m high. Ice formations in the Titlis glacier cave. A glacier cave is a cave formed within the ice of a glacier. Glacier caves are often called ice caves, but the latter term is properly used to describe bedrock caves that contain year-round ice

Question (Q)

How are glacier caves formed?

Reading Comprehension with Attention

Bi-LSTM for Reading Comprehension with Attention



Document (D)

A partly submerged glacier cave on Perito Moreno Glacier. The ice facade is approximately 60 m high. Ice formations in the Titlis glacier cave. A glacier cave is a cave formed within the ice of a glacier. Glacier caves are often called ice caves, but the latter term is properly used to describe bedrock caves that contain year-round ice

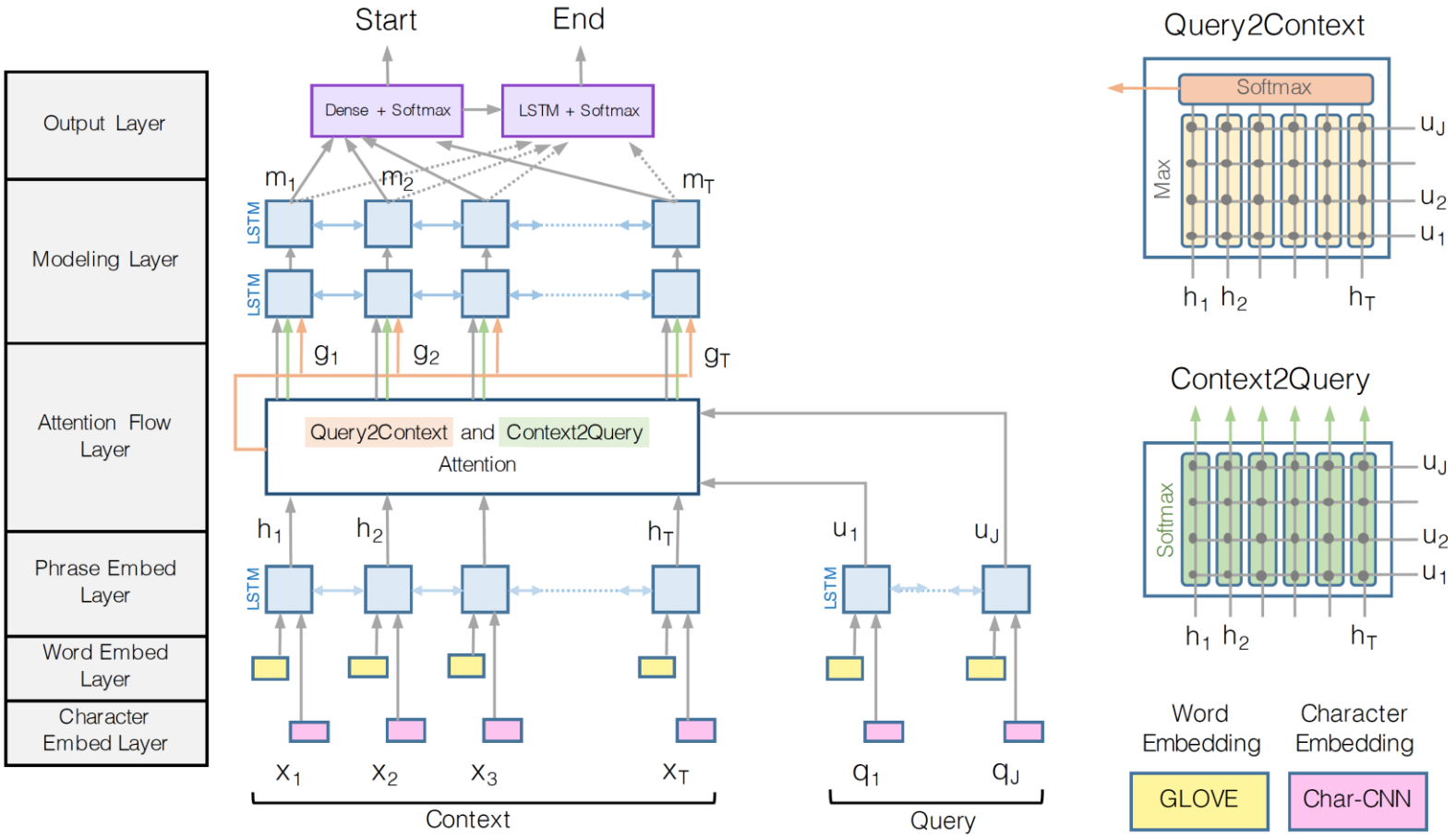
Question (Q)

How are glacier caves formed?

Reading Comprehension with Attention

Bi-Directional Attention Flow (Bi-DAF)

Bi-Directional Attention Flow for Machine Comprehension *(Seo et al. 2017)*



Reading Comprehension with Attention

Bi-Directional Attention Flow (Bi-DAF)

Attention Flow layer is the core idea!

- ***Variants and improvements to the Bi-DAF architecture over the years***

Attention should flow both ways:

1) the context → the question (C2Q)

2) the question → the context (Q2C)

Both attentions are derived from a ***shared similarity matrix*** *between the* ***context (H) and the query (U),*** where S_{tj} indicates the similarity between t-th context word and j-th query word

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

Reading Comprehension with Attention

Bi-Directional Attention Flow (Bi-DAF)

Attention Flow layer is the core idea!

- ***Variants and improvements to the Bi-DAF architecture over the years***

Attention should flow both ways:

1) the context → the question (C2Q)

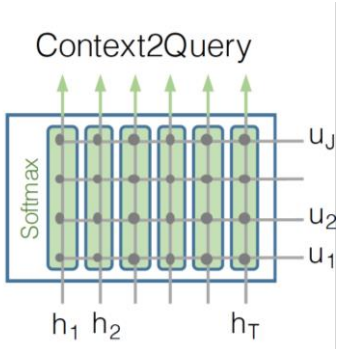
2) the question → the context (Q2C)

1. Context-to-Question (C2Q) attention:

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

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$



Reading Comprehension with Attention

Bi-Directional Attention Flow (Bi-DAF)

Attention Flow layer is the core idea!

- ***Variants and improvements to the Bi-DAF architecture over the years***

Attention should flow both ways:

1) the context → the question (C2Q)

2) the question → the context (Q2C)

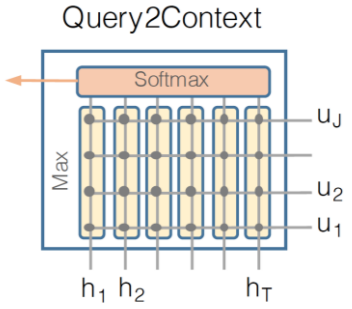
2. Question-to-Context (Q2C) attention:

- *the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max*

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \text{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$



Reading Comprehension with Attention

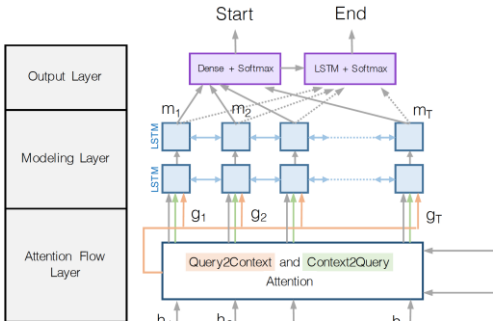
Bi-Directional Attention Flow (Bi-DAF)

A "modelling" layer:

- *Another deep (2-layer) Bi-LSTM over the passage*

And answer span selection is more complex:

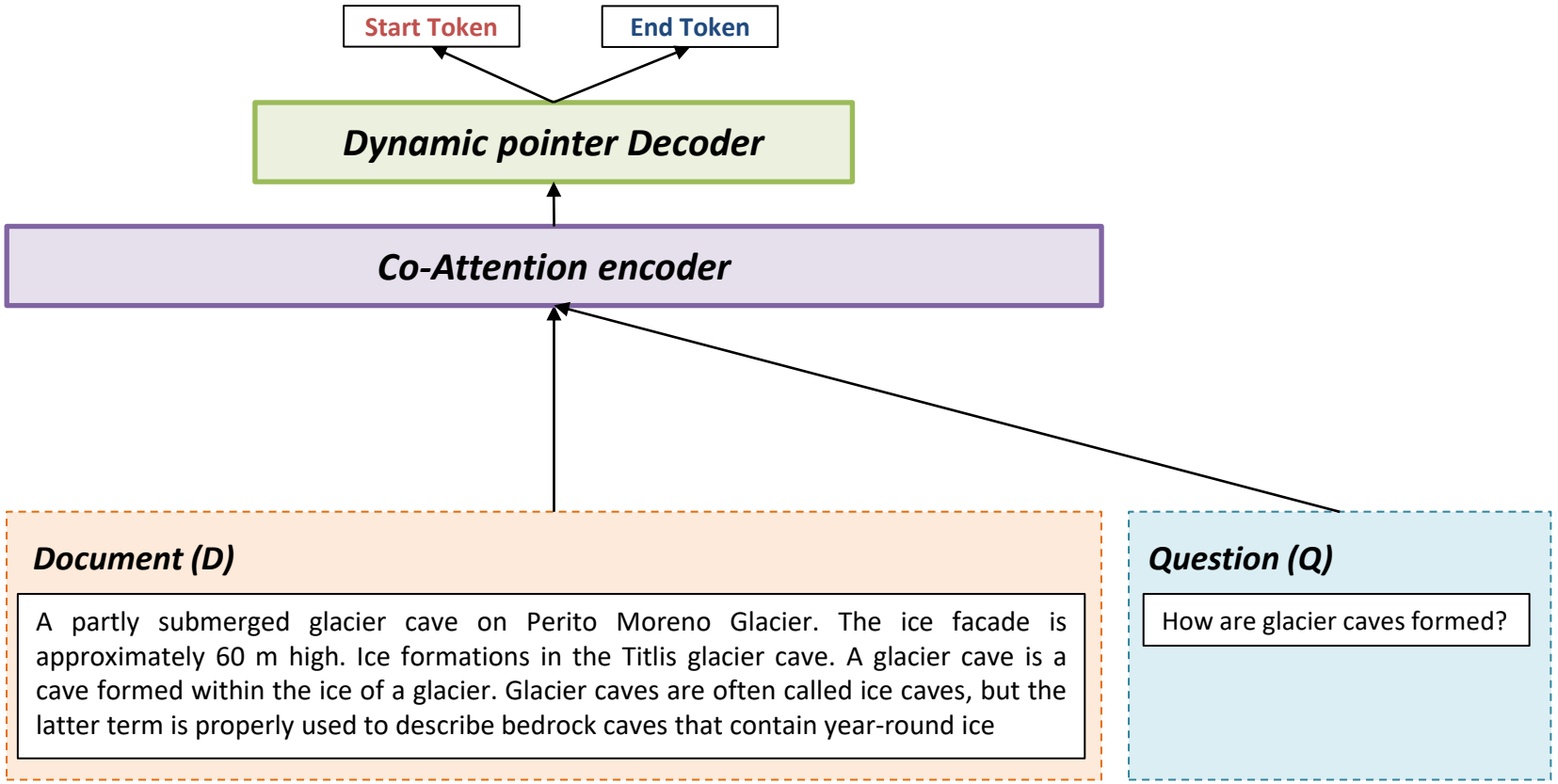
- ***Start**: Pass output of BiDAF and modelling layer concatenated to a dense FF layer and then a softmax*
- ***End**: Put output of modelling layer M through another BiLSTM to give M_2 and then concatenate with BiDAF layer and again put through dense FF layer and a softmax*



Reading Comprehension with Attention

Dynamic Coattention Networks for Question Answering (Xiong 2017)

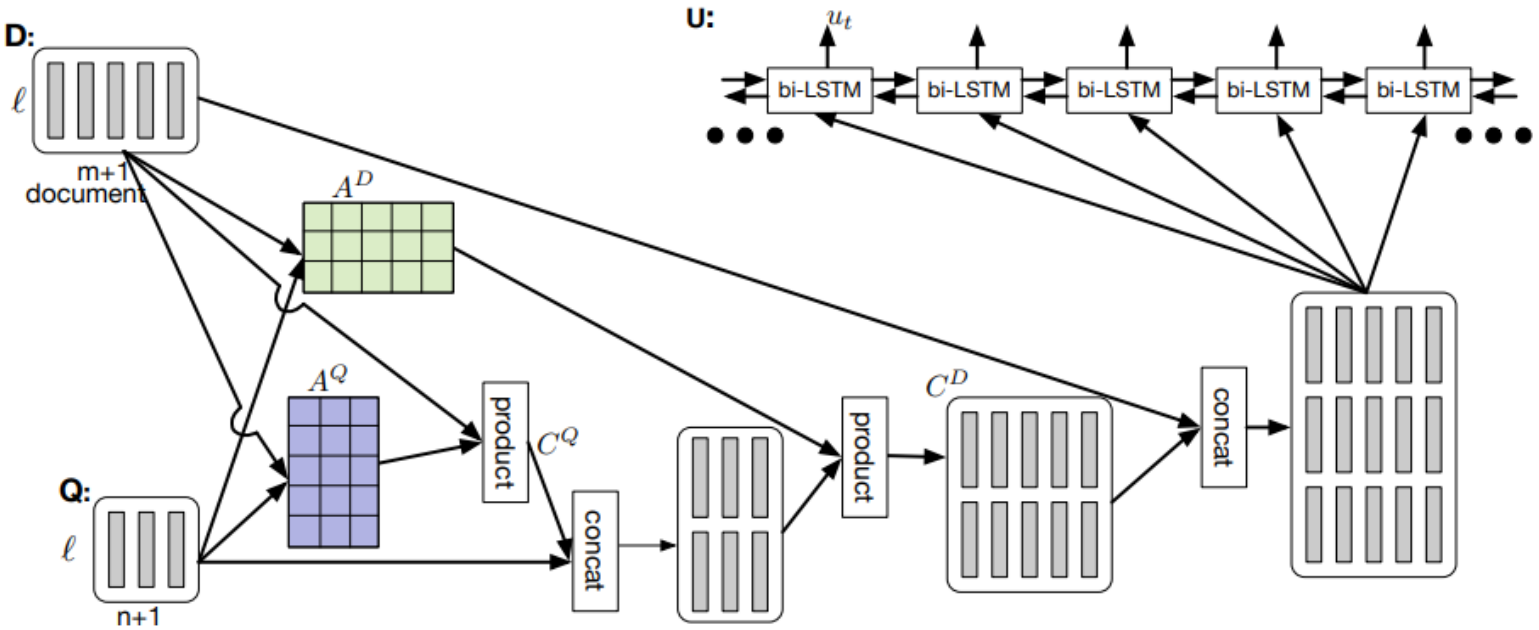
Coattention provides a two-way attention between the context and the question.



Reading Comprehension with Attention

Dynamic Coattention Networks for Question Answering (Xiong 2017)

- *Coattention layer again provides a two-way attention between the context and the question*
- *Coattention involves a second-level attention computation:*
  - *attending over representations that are themselves attention outputs*



More…?

More Advanced Architecture? Preview for next week

The transformer, based solely on attention mechanisms.

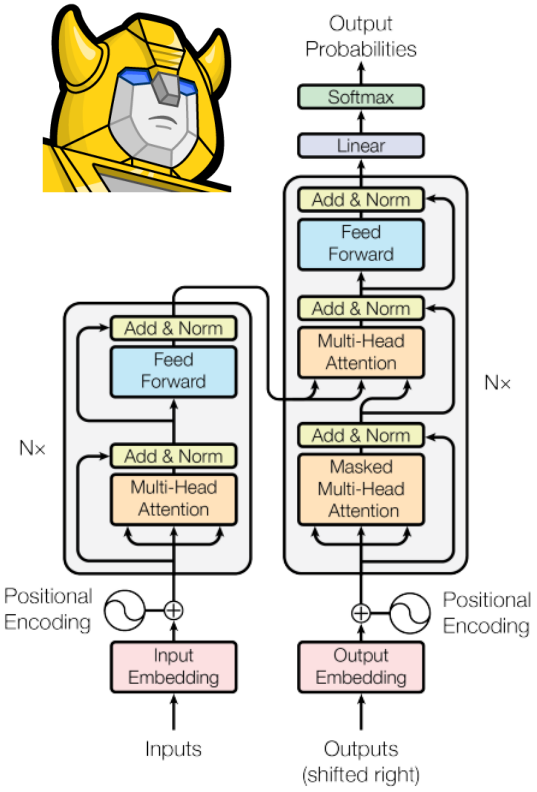
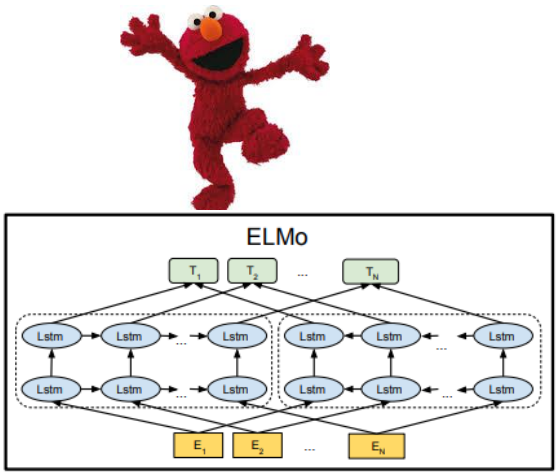
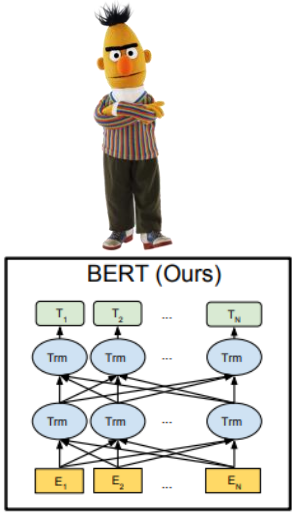


Figure 1: The Transformer - model architecture.

(Vaswani et al, 2017)



(Peters et al, 2018)



(Devlin et al, 2018)

Using Contextual word representations

Question Answering

Research Areas in Question Answering

Research Area	Details
Knowledge-based QA (Semantic Parsing)	• Answer is a logical form, possible executed against a Knowledge Base • Context is a Knowledge Base
Information Retrieval-based QA • Answer sentence selection • Reading Comprehension	• Answer is a document, paragraph, sentence • Context is a corpus of documents or a specific document
Visual QA	• **Answer is simple and factual** • **Context is one/multiple image(s)**
Library Reference	• Answer is another question • Context is the structured knowledge available in the library and the librarians view of it.

Visual Question Answering

Textual Question Answering: Recap

Answer questions by exploiting pure natural language.

Document / Passage

Caren watched TV last night. There was a guy playing tennis. Caren did not know who he is. He was wearing white shirts …

Question

What was he doing?

Answer

Playing tennis

Visual Question Answering

Visual QA

Several questions require context outside of pure language.

Question

What was he doing?

Answer

Playing tennis

Visual Question Answering

Visual QA Datasets

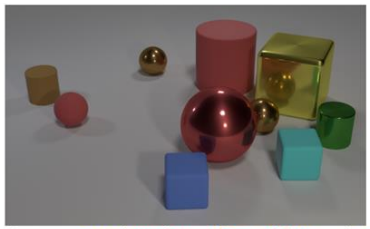
Recently, there are a number of visual QA datasets have sprung up. Some of the more popular ones include:

Type #1: Real images

VQA v2.0 (Goyal et al. 2017)

Type #2: Semantic reasoning

CLEVER (Johnson et al. 2016)



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Type #3: Combined

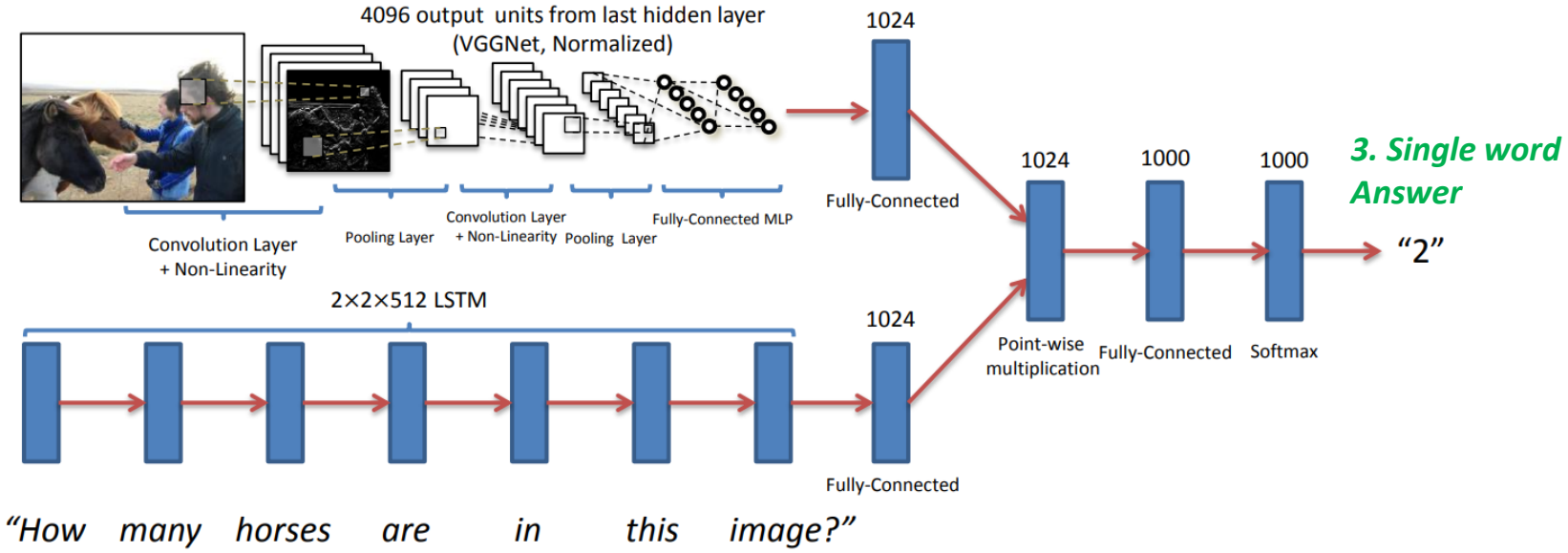
GQA (Hudson and Manning, 2019)

*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

Visual Question Answering

How does it work?

2. Context is a single picture using convolutional neural network (CNN)



3. Single word Answer

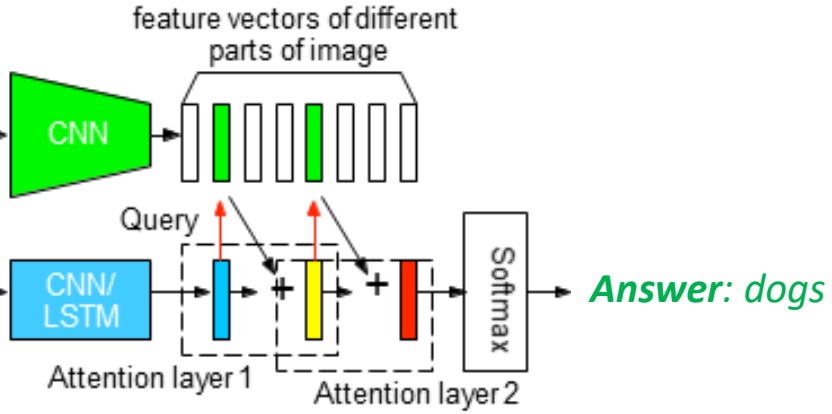
1. Encode sentence with sequence models

Visual Question Answering

How does it work?

*The idea of Visual QA is exactly same as reading comprehension-oriented. Why can't we use **Attention** then? (Yang et al. 2015)*

***Question**: What are sitting in the basket on a bicycle?*



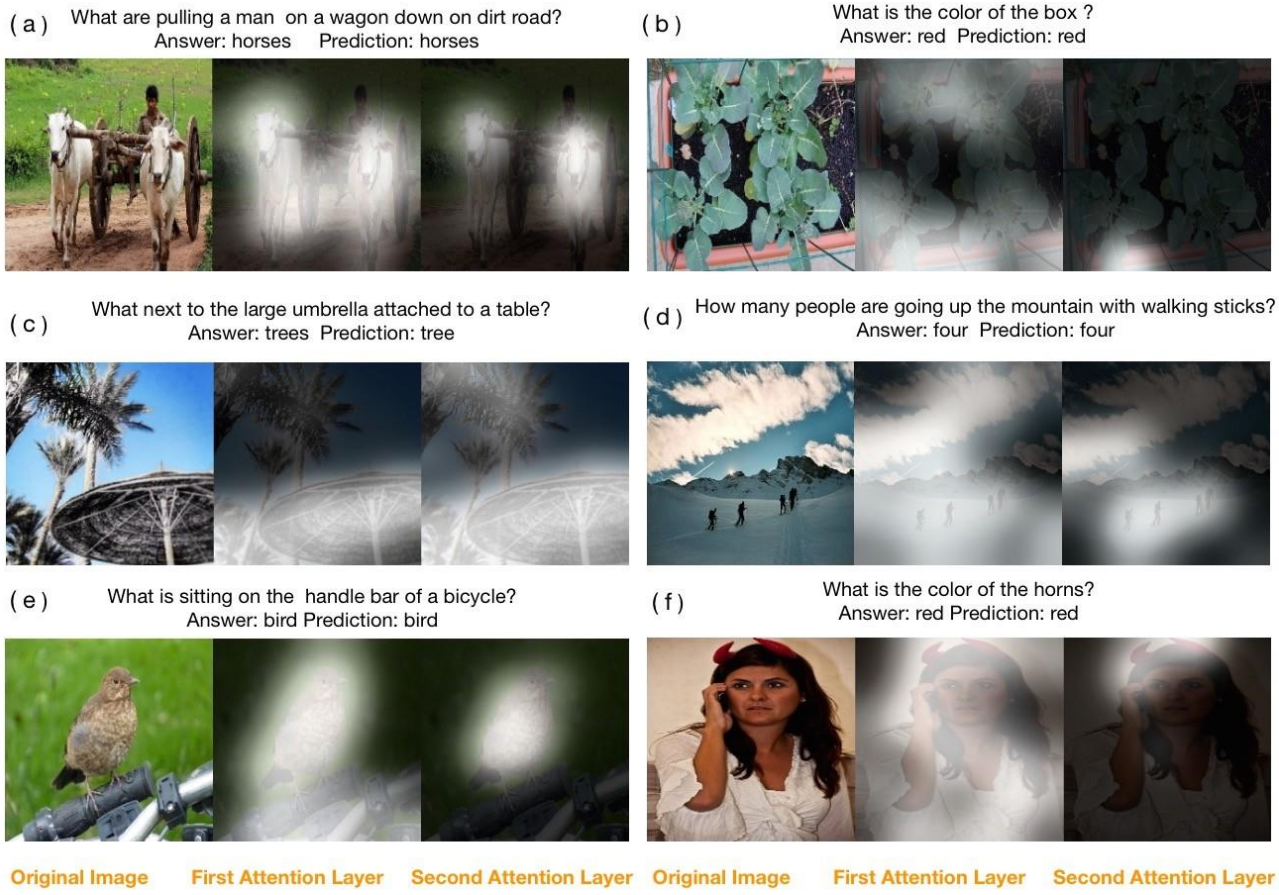
***Answer**: dogs*

Yang et al. (2015): Stacked Attention Networks for Image Question Answering

Visual Question Answering

Visual QA with Attention

Let's try some example. VisualQA (http://vqa.cloudcv.org/)



Visual Question Answering

Visual QA with Attention (Usyd NLP Group 2020)

Question

Where is the Dog Sleeping?

Text to Scene Graph

Image

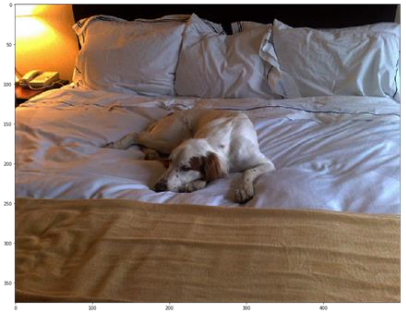
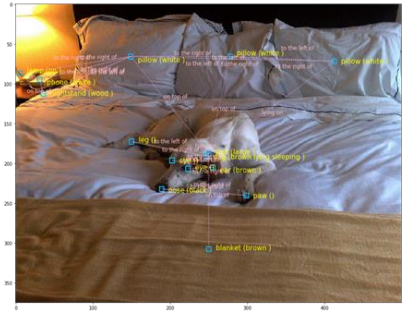
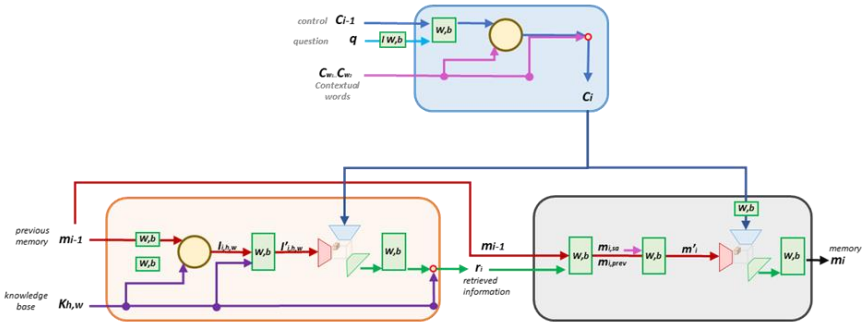


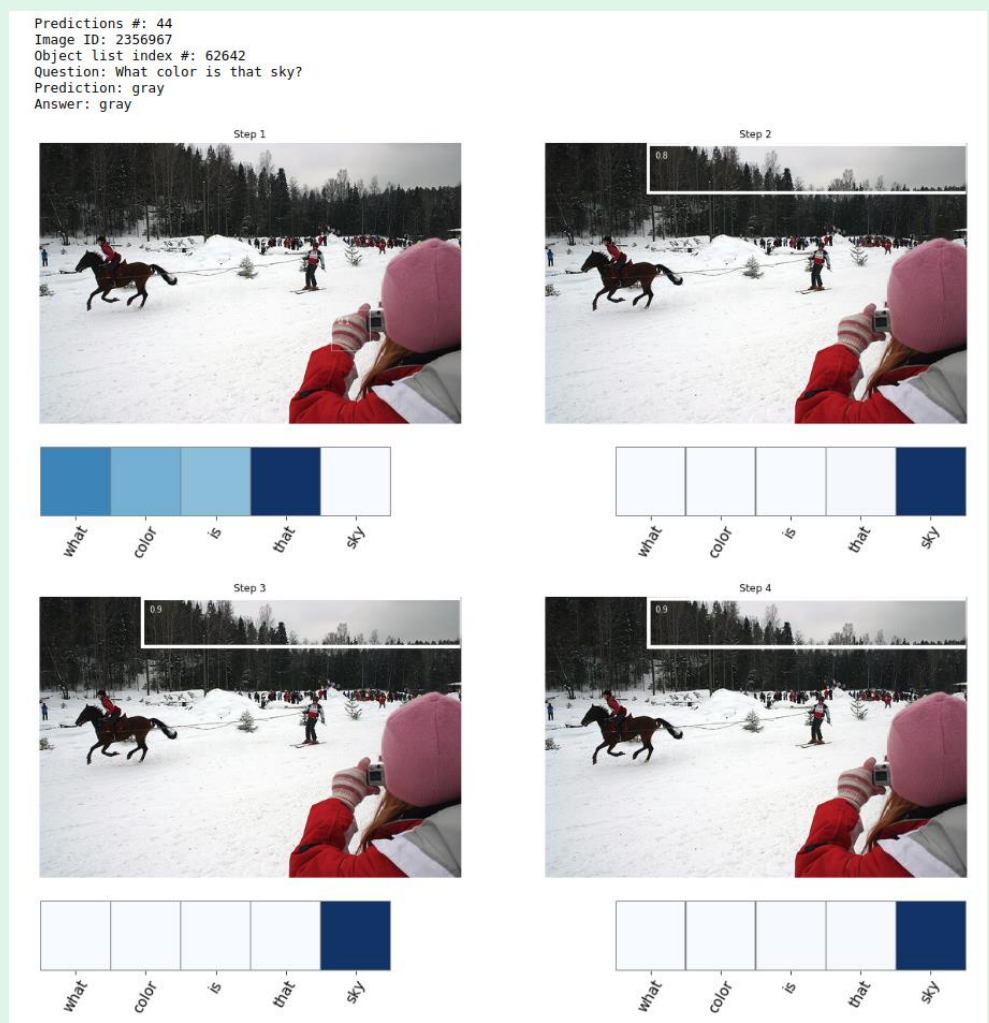
Image to Scene Graph





Visual Question Answering

Visual QA with Attention (Usyd NLP Group 2020)

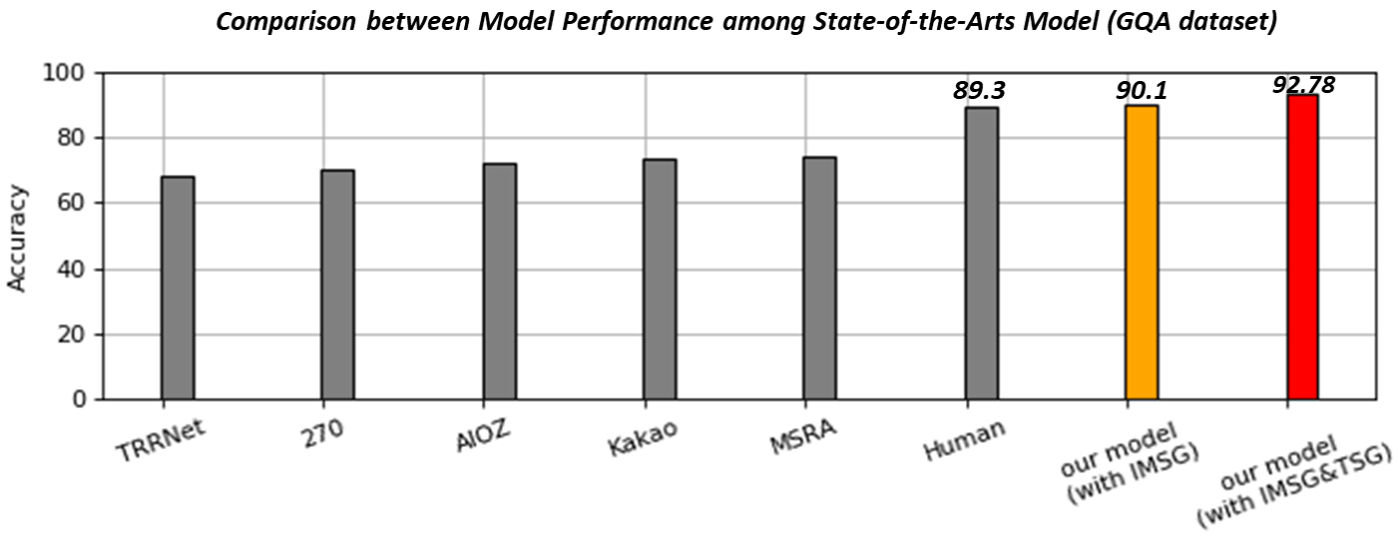


Let's have a look at the demo!!

Visual Question Answering

Visual QA with Attention (Usyd NLP Group 2020)

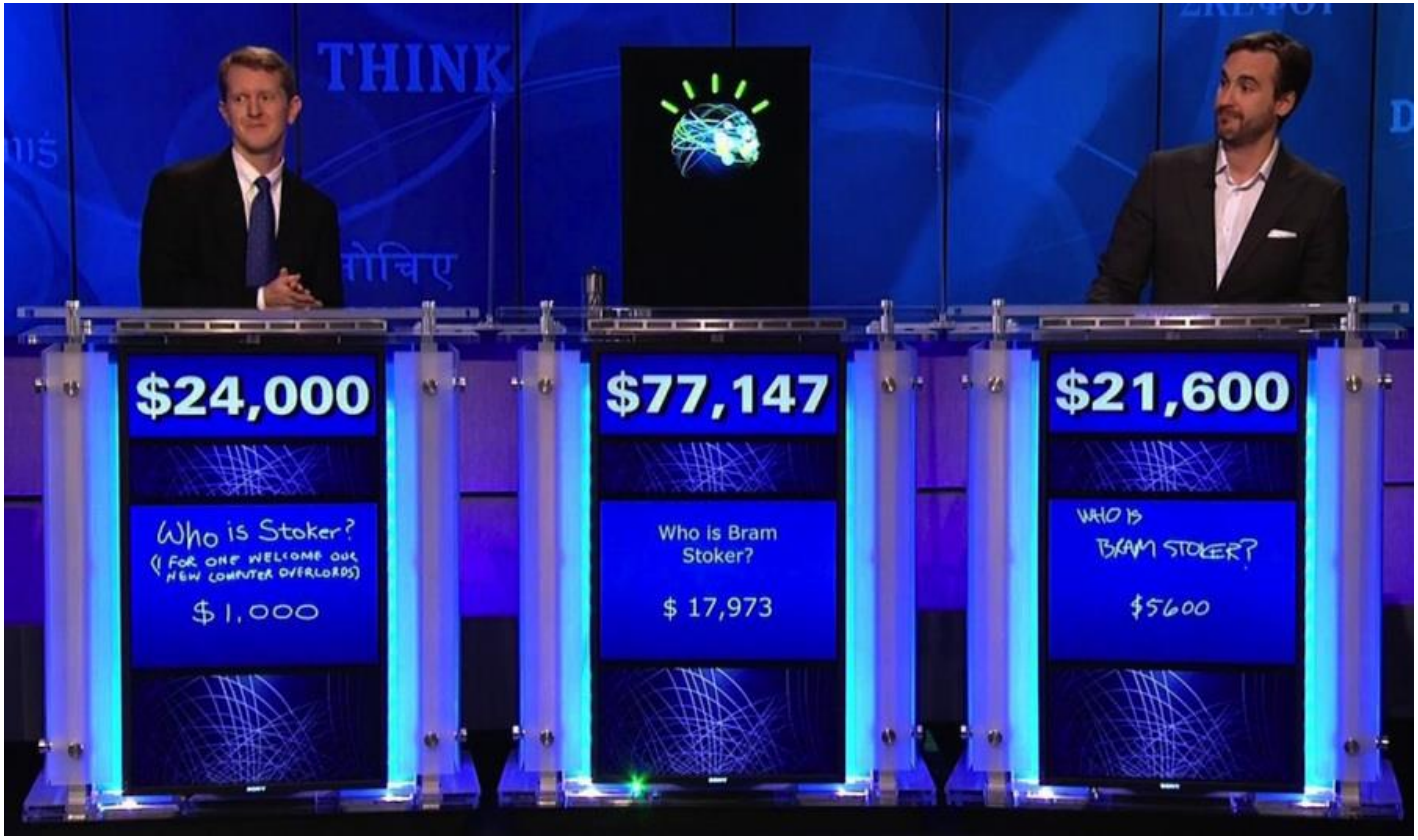
Testing Result



Additional QA

There is no reason to limit to just IR-based or Knowledge-based

Using multiple information sources? IBM Watson!

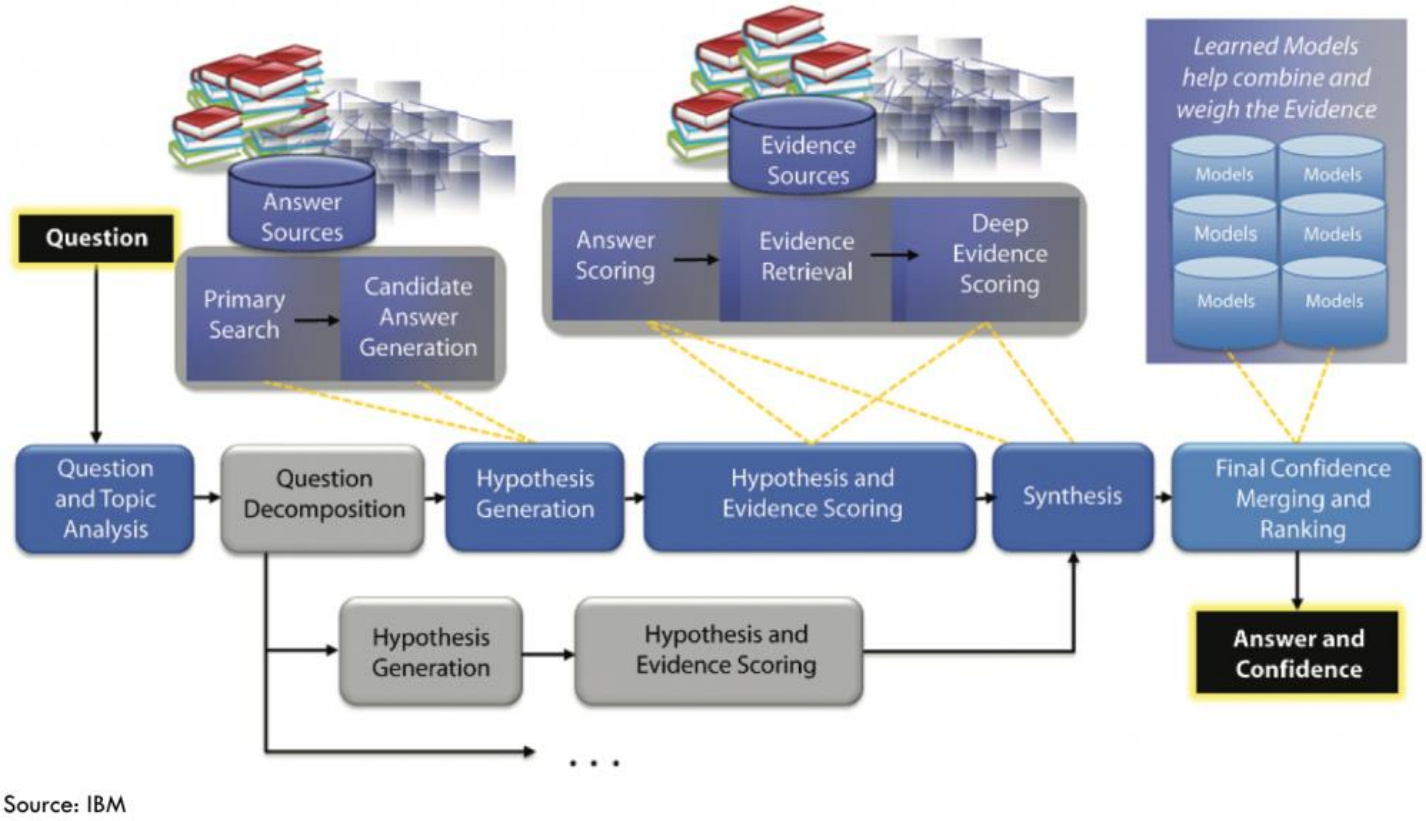


https://youtu.be/P18EdAKuC1U

Additional QA

Using multiple information source

IBM Watson!



https://youtu.be/DywO4zksfXw

The big picture of NLP

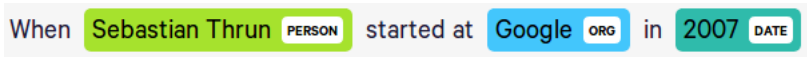
The purpose of Natural Language Processing: Overview

Application

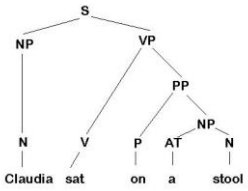
- ***Understanding***
- ***Searching***
- ***Dialog***
- ***Translation***
- ***Sentiment Analysis***
- ***Topic Classification***
- ***Topic Modelling***
- ***....***
- ***Search***
- ***....***

NLP Stack

- ***Entity Extraction*** — When Sebastian Thrun …



- ***Parsing*** — Claudia sat on a stool



- ***PoS Tagging*** — She sells seashells — **[she/PRP] [sells/VBZ] [seashells/NNS]**
- ***Stemming*** — Drinking, Drank, Drunk — **Drink**
- ***Tokenisation*** — How is the weather today — **[How] [is] [the] [weather] [today]**

Reference

Reference for this lecture

- Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.
- Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. " O'Reilly Media, Inc.".
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Manning, C 2018, Natural Language Processing with Deep Learning, lecture notes, Stanford University

- Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1533-1544).
- Chen, D., Bolton, J., & Manning, C. D. (2016). A thorough examination of the cnn/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858.
- Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.
- Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.
- Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.