Memory Networks

SFU NatLangLab

Jetic Gū

Overview

- Focus: Neural Knowledge Incorporation (in NLP)
- Architecture: Dedicated Memory Component in Neural Model
- Core Ideas:
  1. Background: Seq2Seq, Attention, etc.
  2. Memory Network
  3. Applications of Memory Network
  4. Future Work

P1
Why?

Review, and Limitations of Seq2Seq

Including Transformer, BERT, etc.

not HW4P

P1
Why?

Think: Seq2Seq Models

- Powerful
  - Variable-length input: variable-length output
  - Variants: Attentional Seq2Seq, Transformer, etc.
- Uses encoder decoder architecture
  - Encoder: performs feature extraction
  - Decoder: generates output based on aggregated features
- Roll of neural network units: projection functions, from one feature space to another

Review

P1
Why?

Think: Seq2Seq Models

- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)

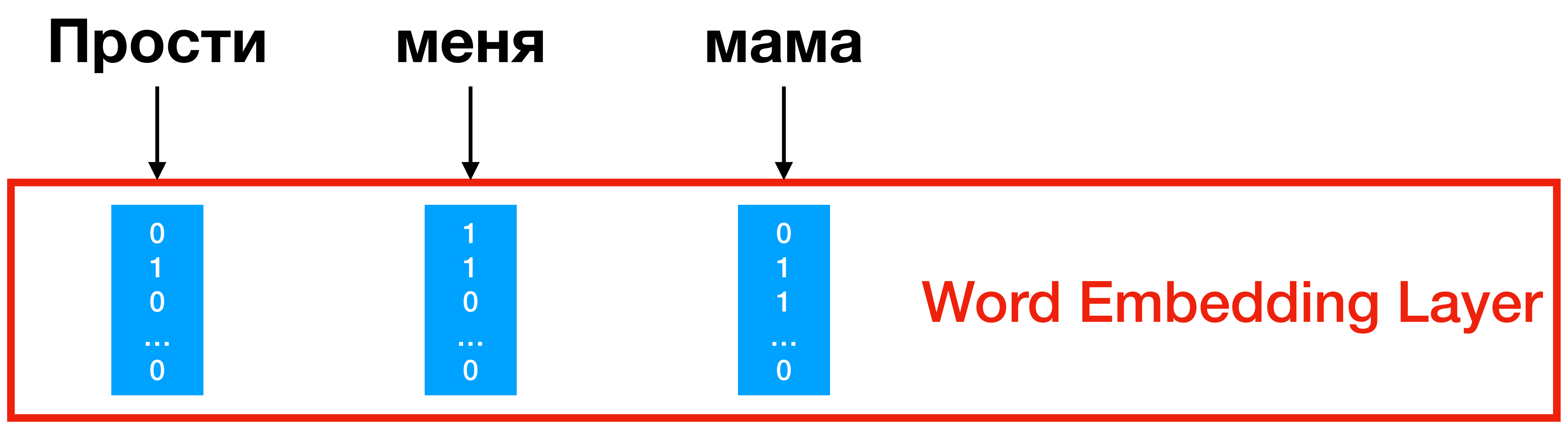
Прости **меня** **мама**

Review

P1
Why?

Think: Seq2Seq Models

- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)

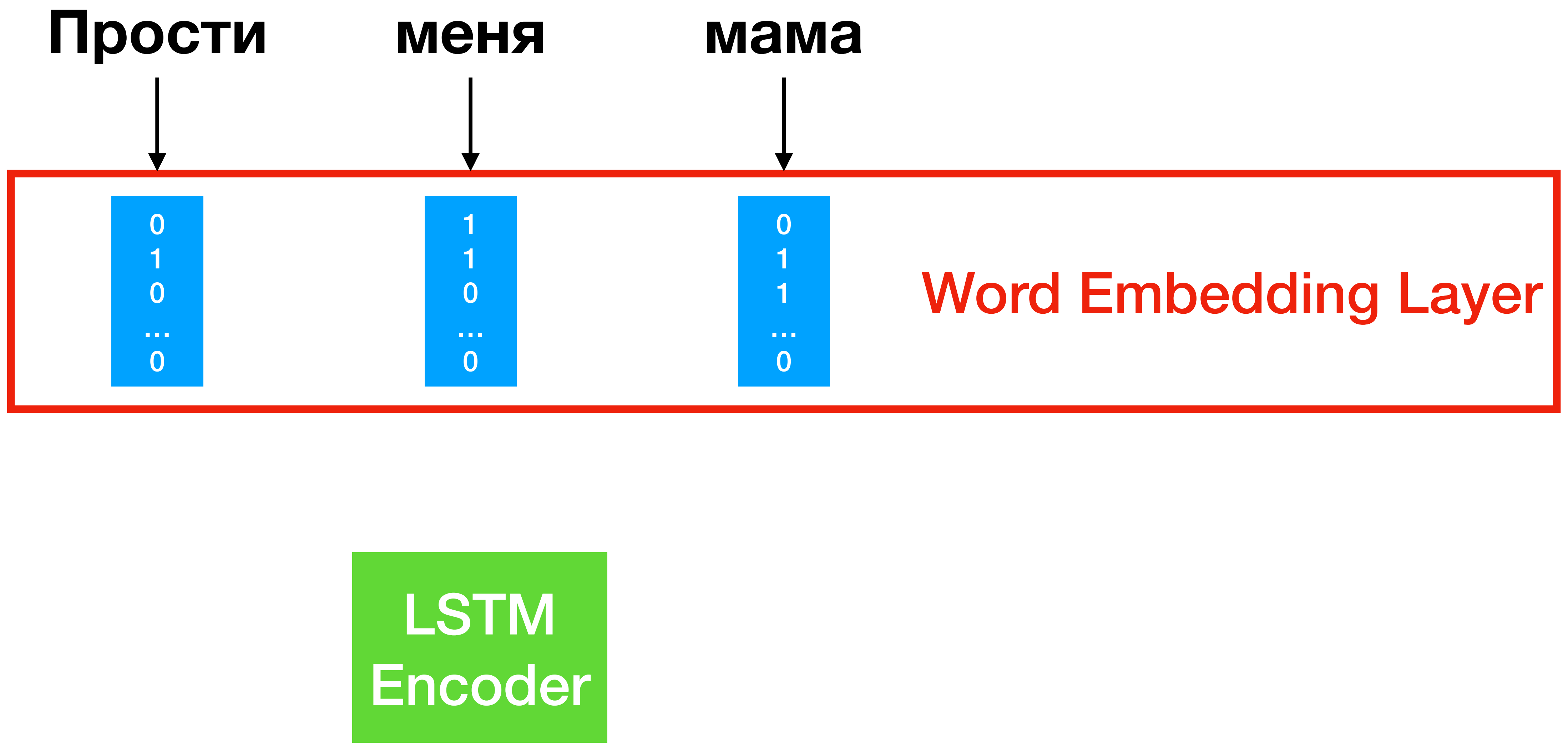


Review

P1
Why?

Think: Seq2Seq Models

- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)

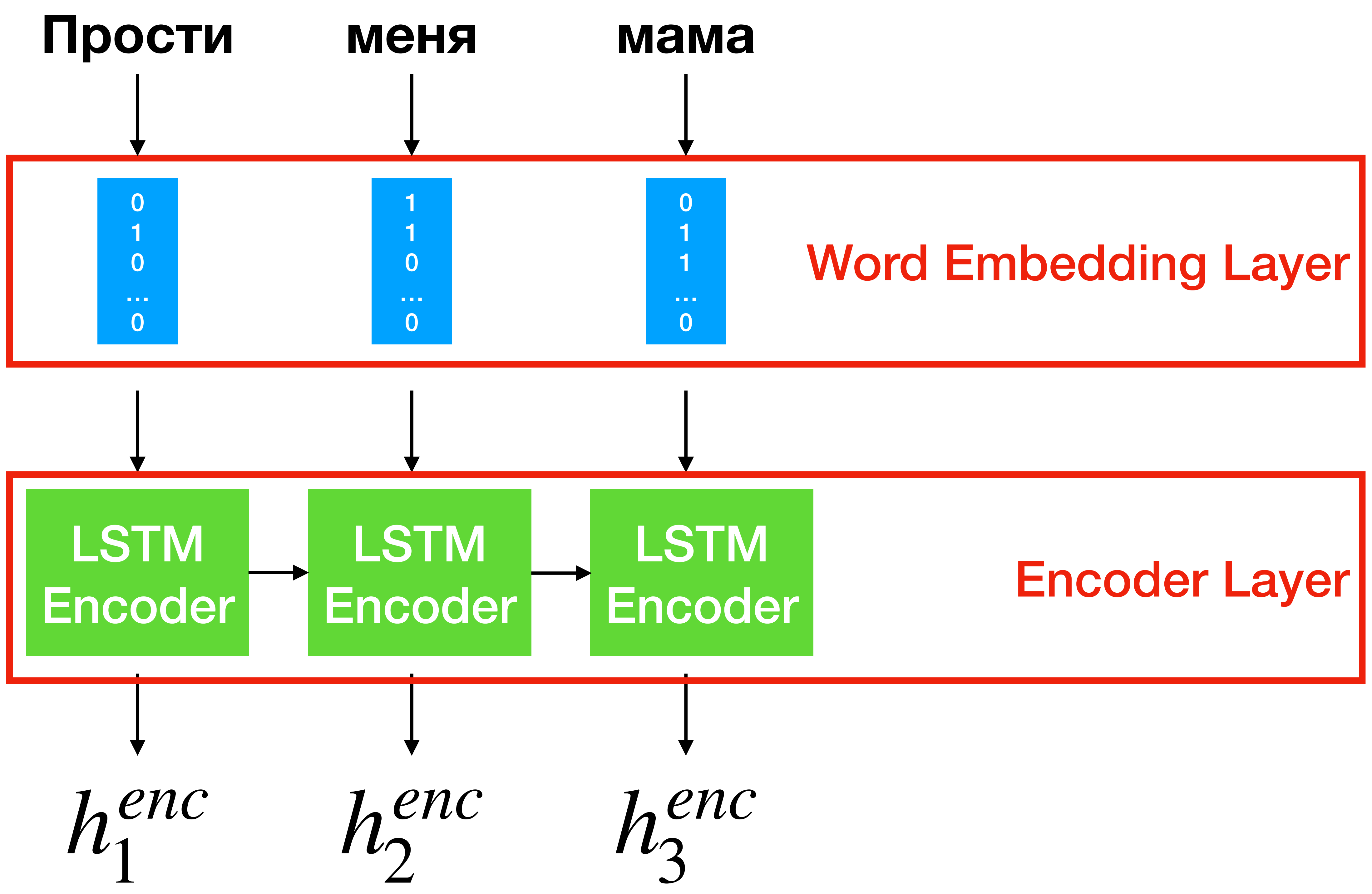


Review

P1
Why?

Think: Seq2Seq Models

- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)

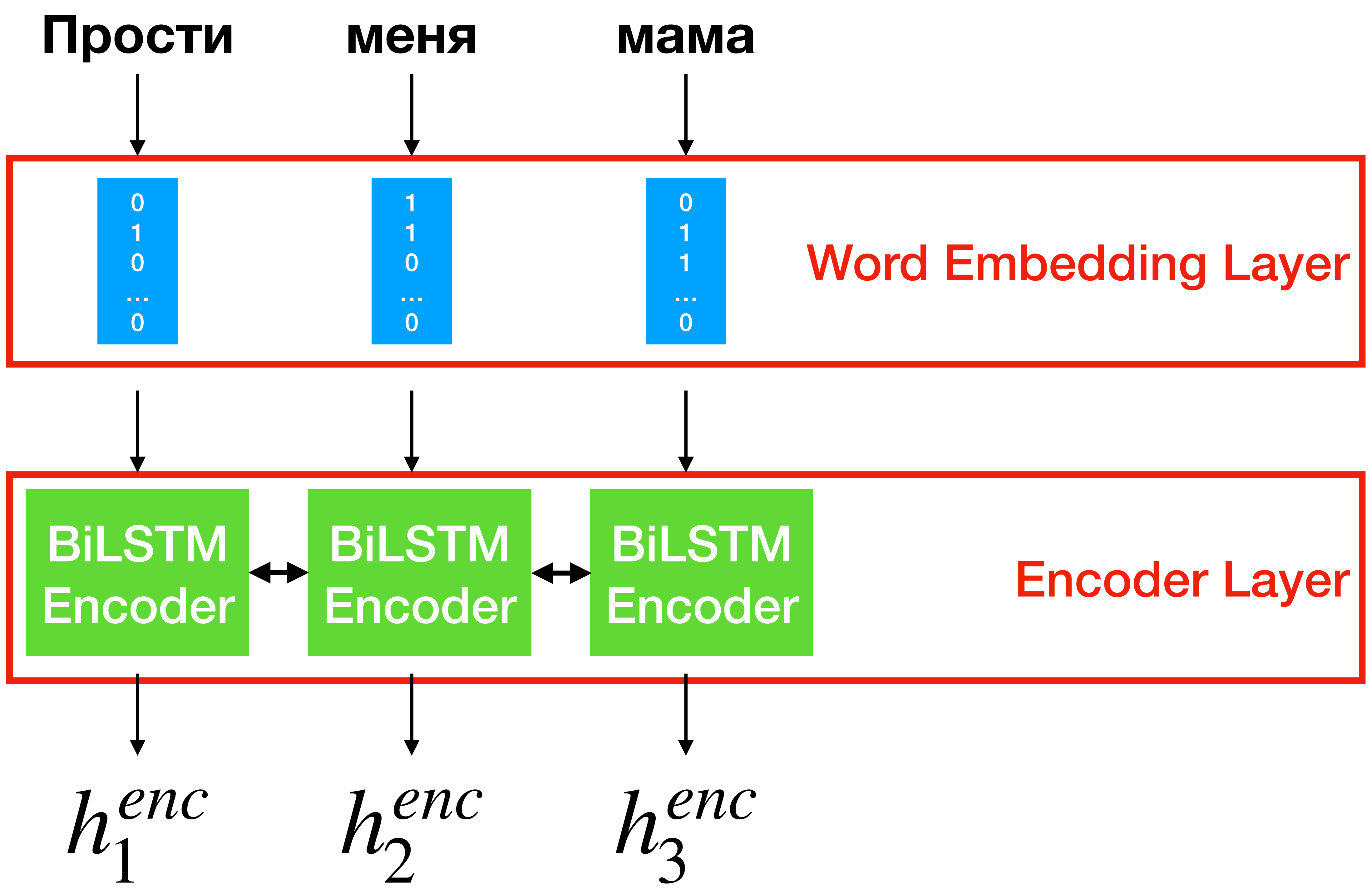


Review

P1
Why?

Think: Seq2Seq Models

- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)

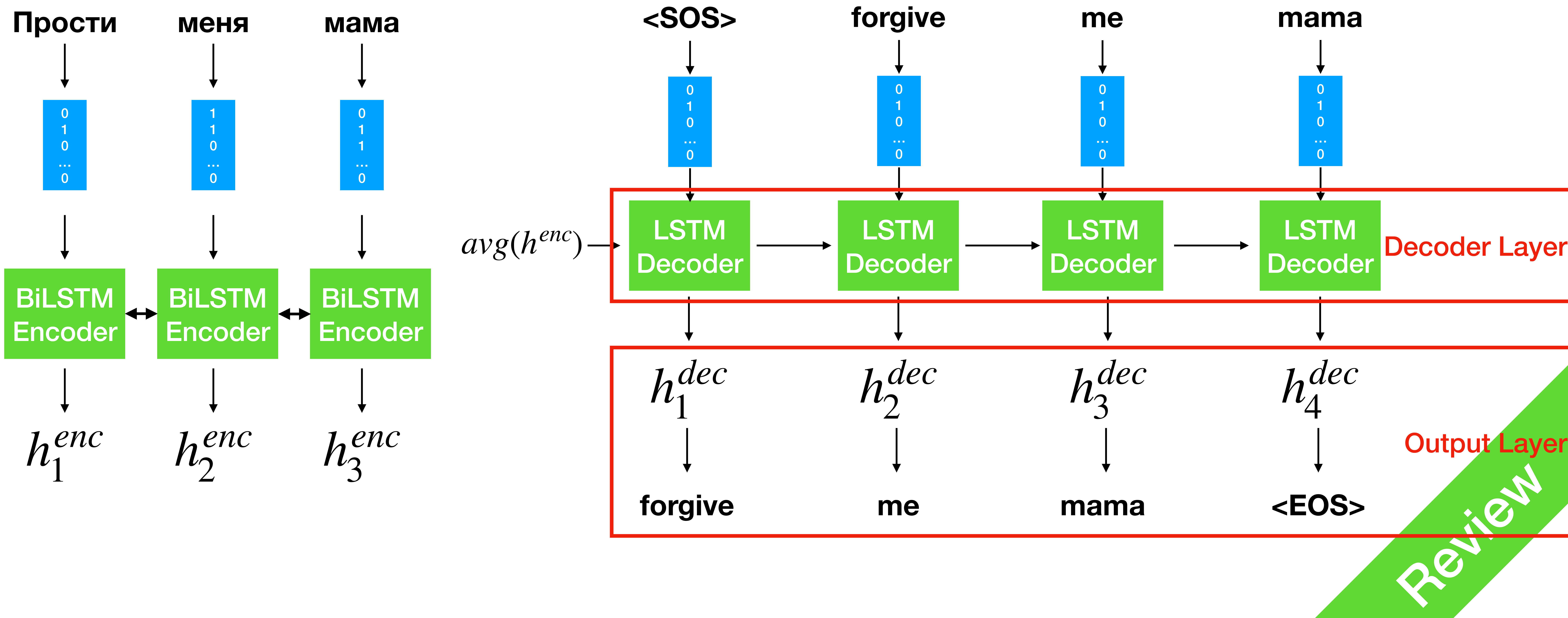


Review

P1
Why?

Think: Seq2Seq Models

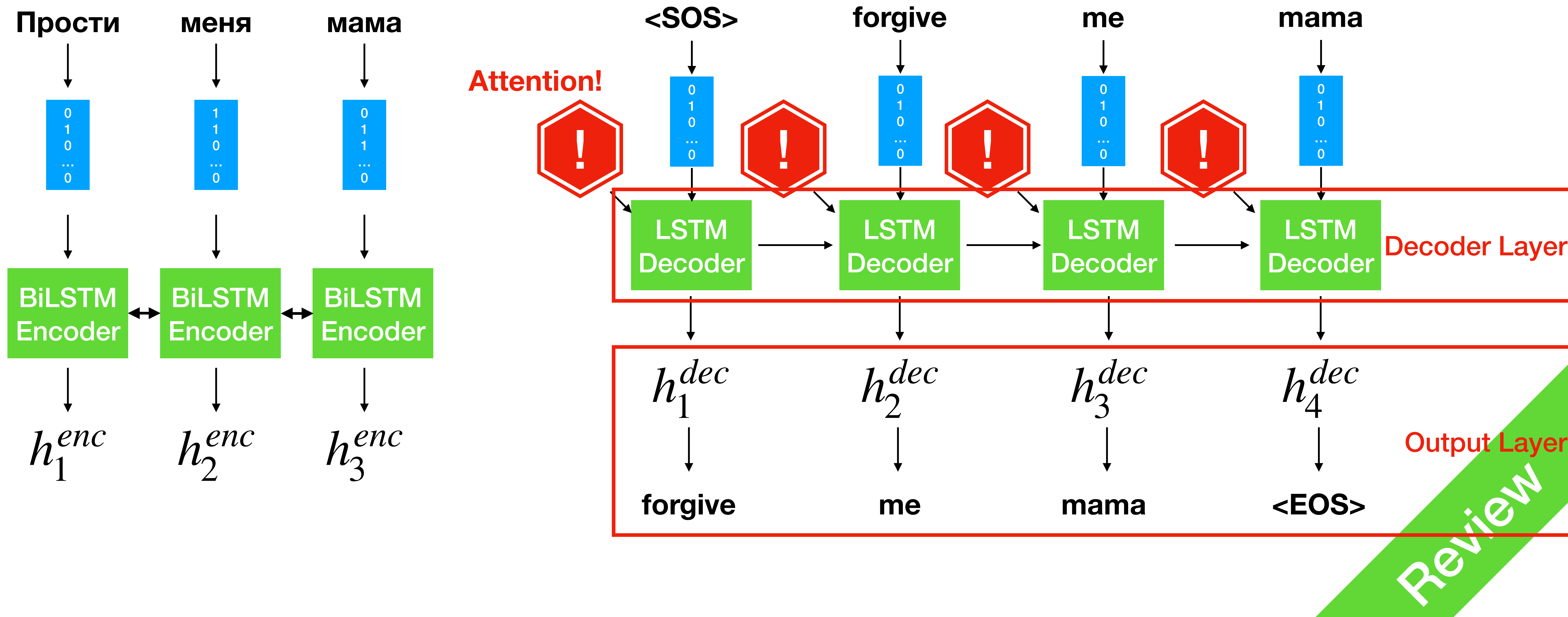
- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)



P1
Why?

Think: Seq2Seq Models

- As an example, LSTM-based Seq2Seq for Translation
  Русский (RU) to English (EN)



P1
Why?

Why does attention work?

- The neural decoder is a conditional language model
  - $P(e_i \mid e_{<i}, F) \in [0,1]^{DictSize}$
  - Initialised with $avg(h^{enc})$: provide condition as a guide
    - $P(e_i \mid e_{<i}, avg(h^{enc})) \in [0,1]^{DictSize}$
    - Problem? Memory!
      Fixed dimension $|h^{dec}|$ for all encoder info and all $e_{<i}$ info

Think

P1
Why?

Why does attention work?

- The neural decoder is a conditional language model
  - $P(e_i \mid e_{<i}, F) \in [0,1]^{DictSize}$
  - Guide with Attention for every step t
    - Project decoder state h_t^{dec} and all encoder states $h_{0:|f|}^{enc}$ into the same feature space, find the most relevant h_i^{enc} and do weighted sum
    - Decoder memory $|h^{dec}|$: ensure fluent language generation
      Encoder memory $|h^{enc}| \times |f|$: ensure src-tgt relevance

Think

P1
Why?

What kind of Knowledge is learned in Seq2Seq for RU-EN?

- Russian Embedding and English Embedding: word-level features

Features

- Encoder BiLSTM: extract useful features from Russian Word Embeddings, word before and after each position

Features Aggregation & Representation Storage

- Decoder LSTM: predict next word given previous English words

Projection

- Attention: project encoder representations and decoder representations, select the most relevant encoder representation for the current time step

Projection

- Output layer: project Attended Decoder representation into target dictionary probability distribution

Projecti

Concept

P1
Why?

Current NLP Approaches

- **Symbolic knowledge** are fed into Neurones (e.g. RNN) for training
- **Limited** in memory **storage capacity**
- **Agnostic to explicit knowledge**, we assume the parameters will pick it up
- Applications: seq2seq, classification, etc.

Review

1. Graves et al., Hybrid computing using a neural network with dynamic external memory, Nature 2016

P1
Why?

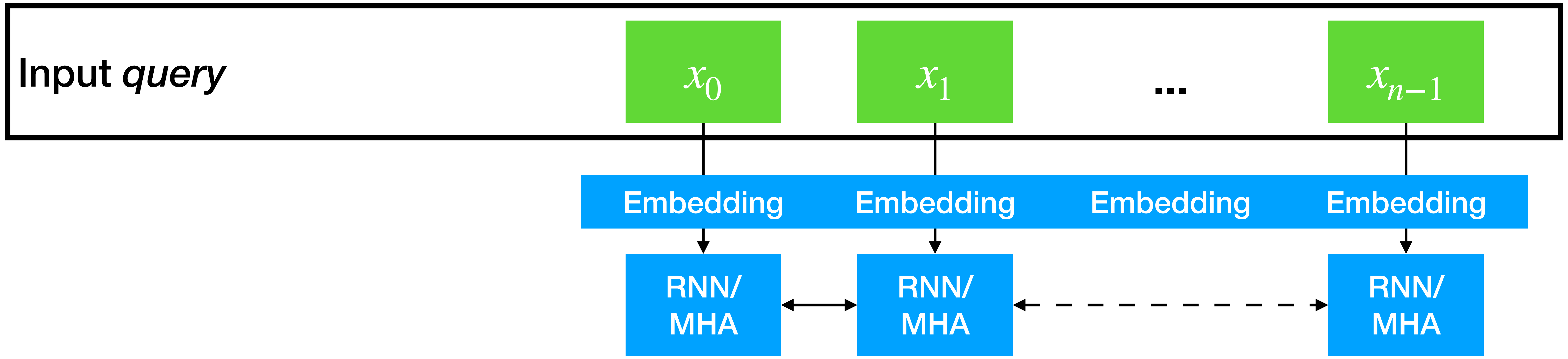
RNN/MHA Units

Input *query* x_0 x_1 ... x_{n-1}

Review

P1
Why?

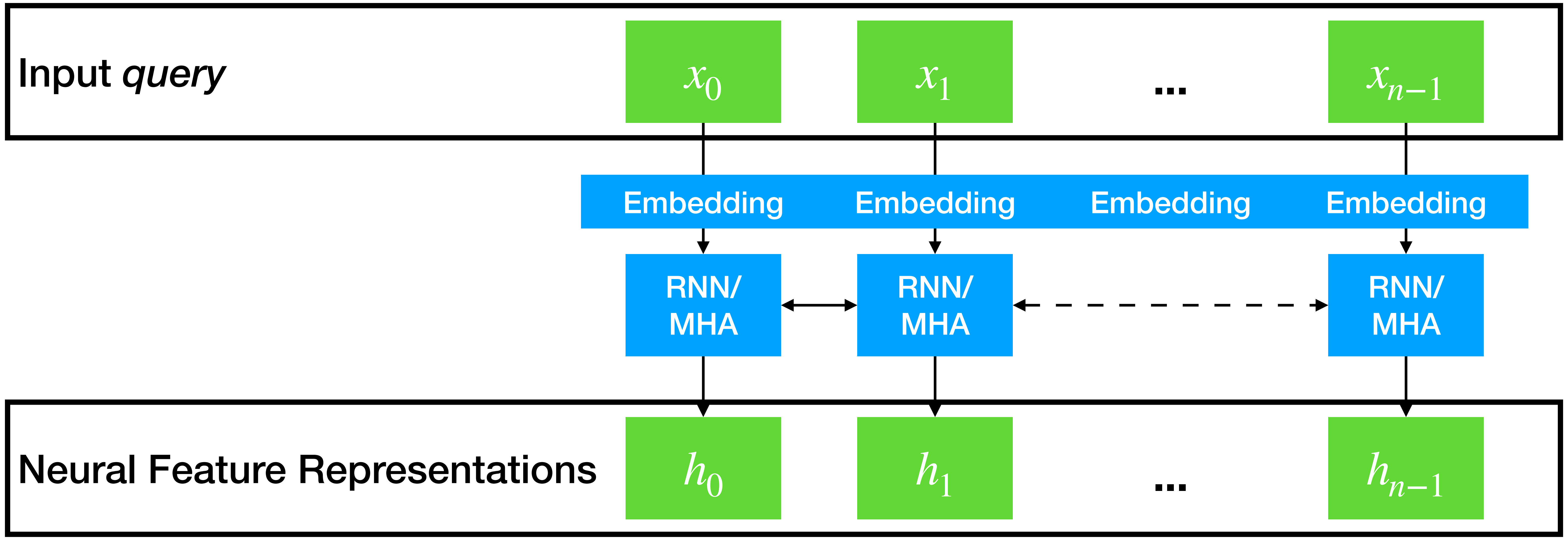
RNN/MHA Units



Review

P1
Why?

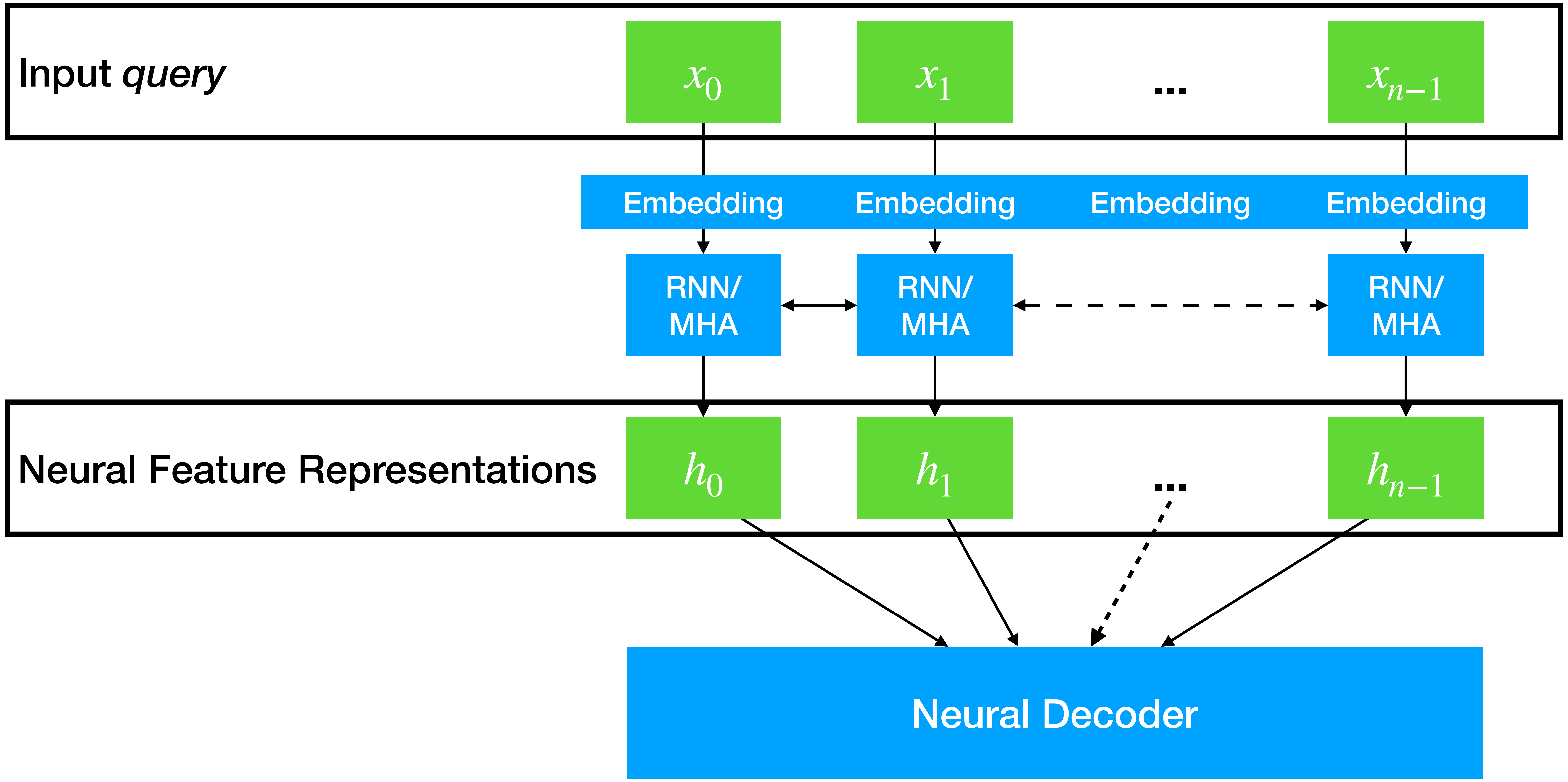
RNN/MHA Units



Review

P1
Why?

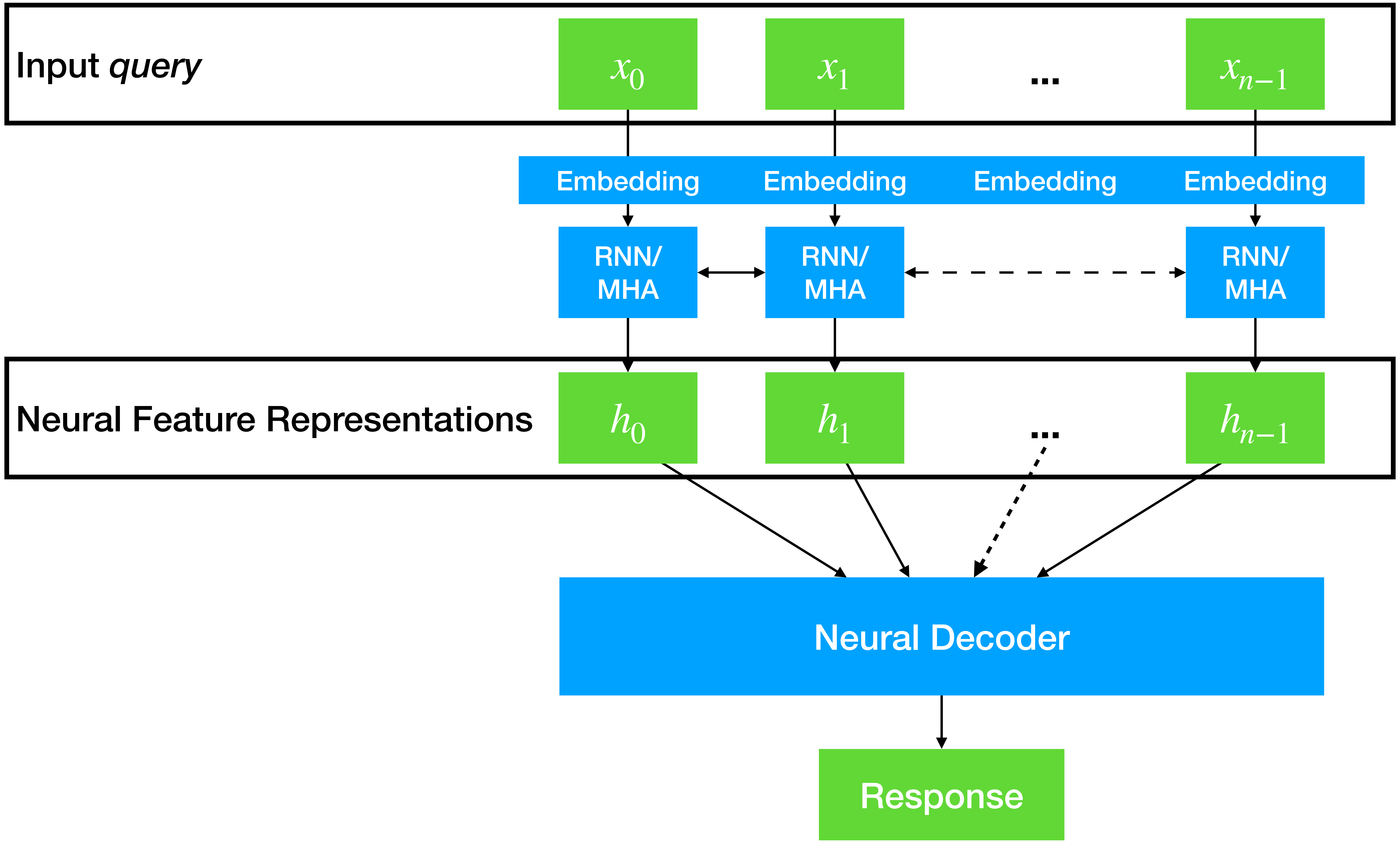
RNN/MHA Units



Review

P1
Why?

RNN/MHA Units

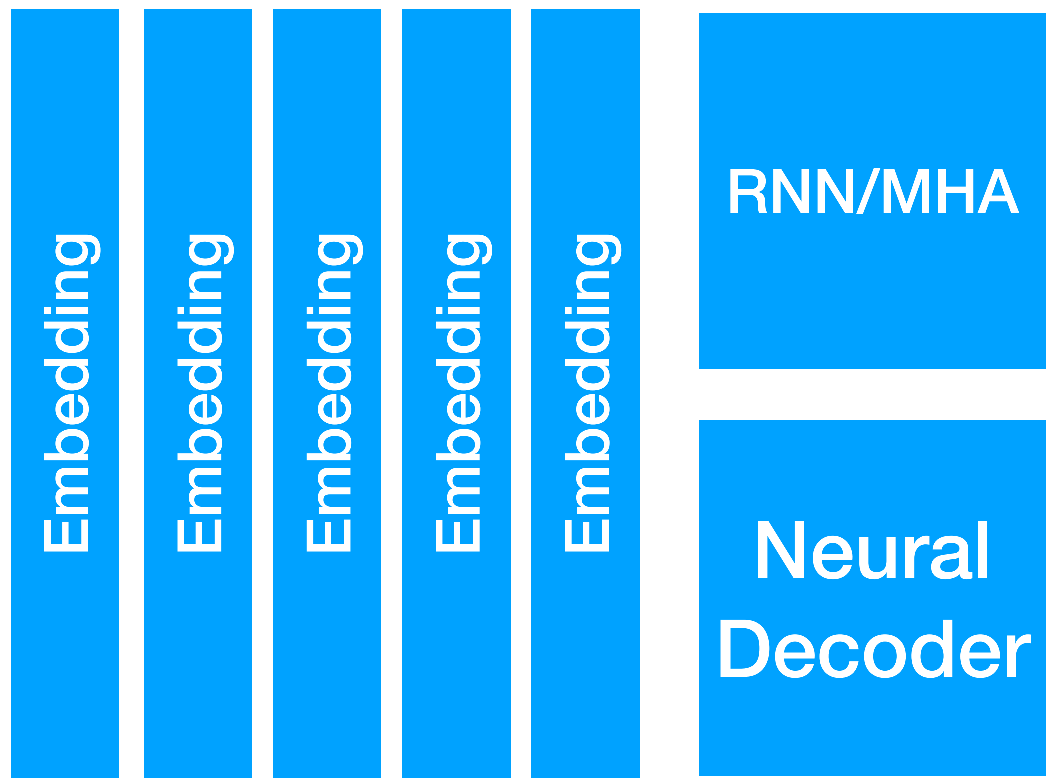


Review

P1
Why?

Current NLP Approaches

- Neural Components
    - *Query*-to-*Response* Mapping Function
- Expect: **limited parameters** learn all knowledge
- Reality: sometimes you need information external to the input: ***Context***

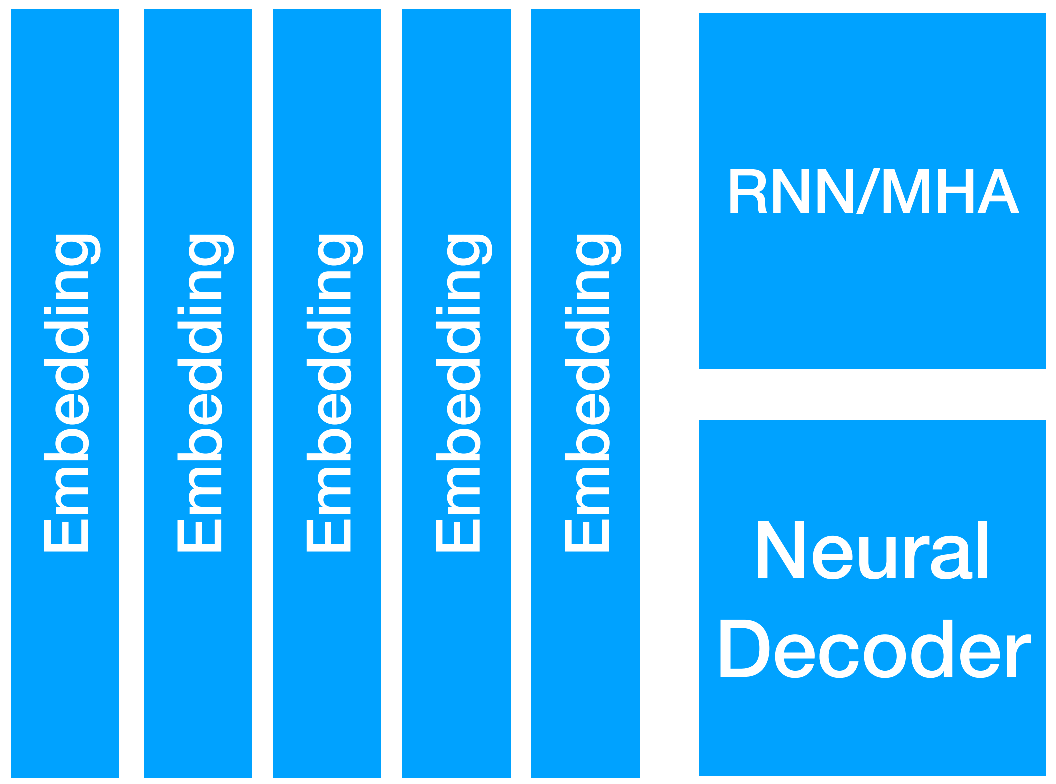


Review

P1
Why?

Current NLP Approaches

- Reading Comprehension
  - Paragraph (Context):
    Fawlty Towers is a British television sitcom written by John Cleese and Connie Booth, broadcast on BBC2 in 1975 and 1979. Two series of six episodes each were made. The series is set in Fawlty Towers, a fictional hotel in the seaside town of Torquay on the English Riviera. **The plots centre on the tense, rude and put-upon owner Basil Fawlty (Cleese), his bossy wife Sybil (Prunella Scales)**, the sensible chambermaid Polly (Booth) who often is the peacemaker and voice of reason, and the hapless and English-challenged Spanish waiter Manuel (Andrew Sachs). They show their attempts to run the hotel amidst farcical situations and an array of demanding and eccentric guests and tradespeople.
  - Query: Who is the actor who played Basil Fawlty in Fawlty Towers?
  - Response: John Cleese



Review

P1
Why?

Using RNN/MHA with *Context*

- Treat external *Context* as part of the *Query*. **Problem:**
  - **Handling Exotic Structures** is difficult
  - **RNN/MHA** has limited **long-term memory capacity**
  - **Complex Internal Dynamics**
    RNN/MHA do not particularly perform very well

!

1. Yih et al., The Value of Semantic Parse Labelling for Knowledge Base Question Answering, InProc ACL2016
2. Dhingra et al., Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access, InProc ACL2017
3. Dong et Lapata, Coarse-to-Fine Decoding for Neural Semantic Parsing, InProc ACL2018

Memory Network Basics

Let’s turn the clock

Remember Me?

Memory Network Definition

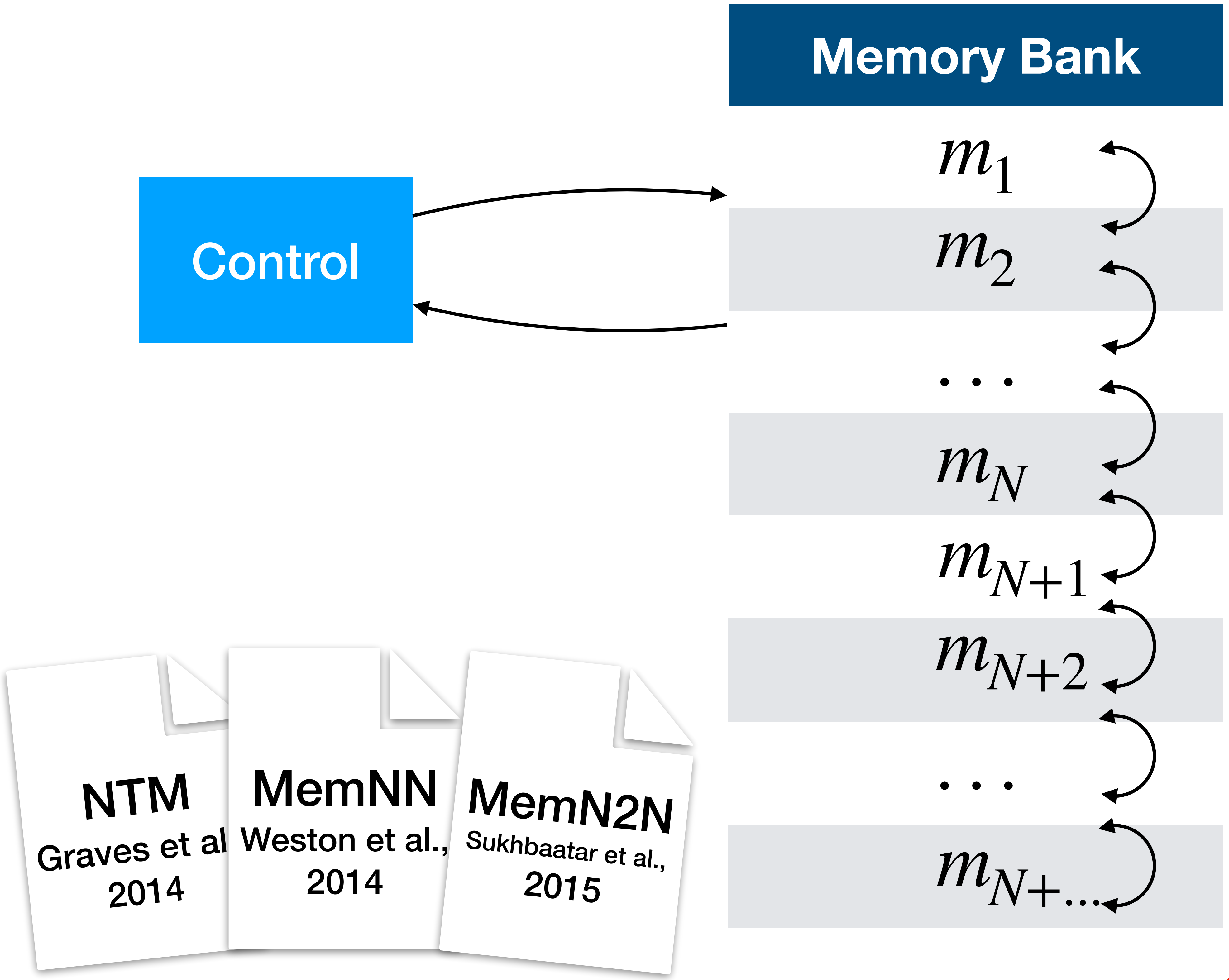
- Definition:
  - A **neural** architecture
  - with **dedicated variable-length neural memory components,**
  - that is capable of **complex internal dynamics**

Memory Bank
m_1
m_2
$\dots$
m_N

Concept

Memory Network Definition

- Core Features
  - **Expandable** Neural Memory Unit
  - **Neural Controller** for Read/Write
  - Complex **Internal Dynamics***



Concept

Memory Network Definition

- Is memory network old?
  - Yes, it goes back to before Bahdanau's RNNSearch
- Why haven't I heard of memory networks?
  - Because people don't often refer to them as Memory Networks (we'll come back to this)

Memory Bank
m_1
m_2
$\dots$
m_N

Concept

P2
Memory Network

Memory Network MemNN

- First End-to-End application in NLP
- Database QA
  - Multiple sentences from a database is given as input ***Context***
  - The model is expected to answer a ***Query***

Memory Content (Plain)
Barack Obama was the president of USA.
Cheddar is the most popular cheese.
NLP is complete nonsense.
....
Monty Python is the greatest comedy group.

Some 1k facts

Query:
What is Monty Python?

Response:
One of the greatest comedy groups.

Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

Memory Network
MemNN

- Storage Structure:
  - Each slot stores one encoded sentence
    - LSTM-based, or BERT e.g.
  - Once written, the representation doesn't receive update

Memory Bank	Memory Content (Plain)
m_1	Joe is in the Kitchen.
m_2	Joe is with Jack.
$\dots$	...
m_N	They go to the theatre together

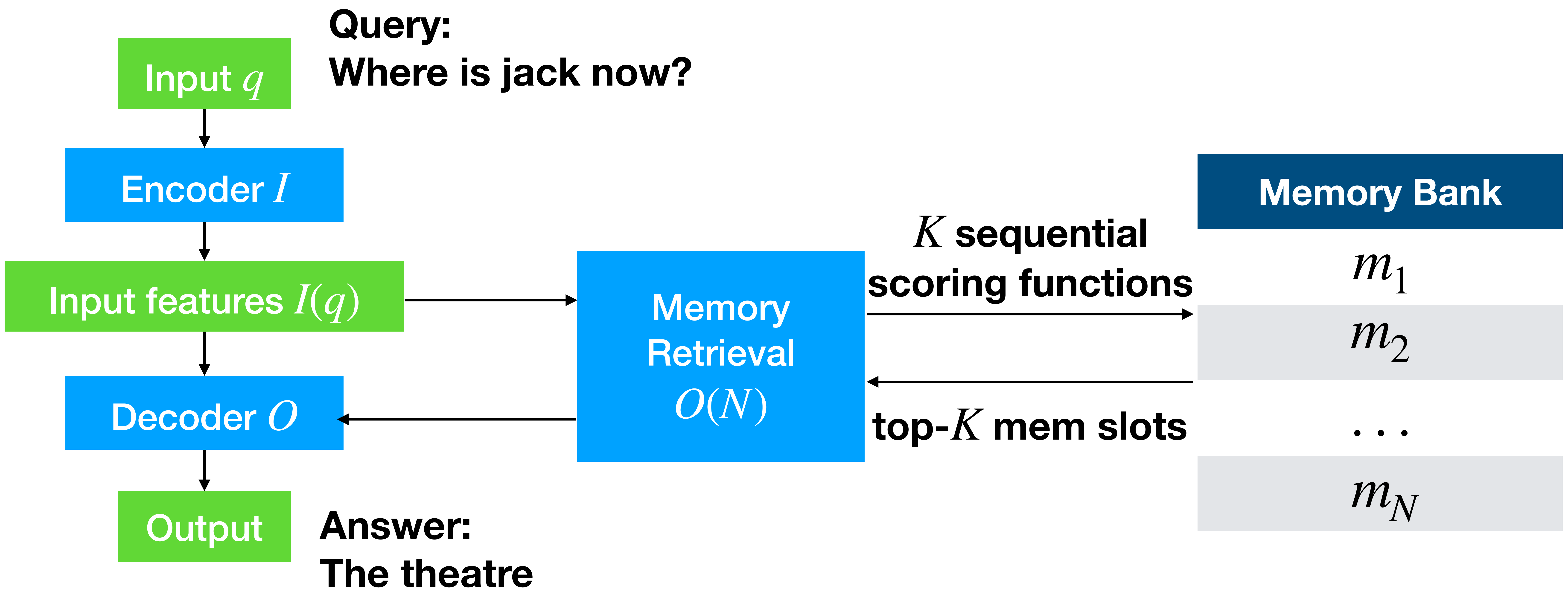
Query:
Where is jack now?

Answer:
The theatre

Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

Memory Network
MemNN



1. Weston et al., Memory Networks, InProc ICLR 2015
2. Sukhbaatar et al., End-To-End Memory Networks, InProc NIPS 2015

Detail

P2
Memory Network

MemNN Response

- Decoding module O
  - Selects k supporting memory cells $m_{o_1}, \ldots, m_{o_k}$
    - $o_k = O(x, \mathbf{M}) = \underset{i=0,\ldots n-1}{\text{argmax}}\; s_o([x, o_{<k}], m_i])$[2]
  - Can also use Attention Mechanism instead
    - $o = \sum_i w_i m_i$
- In Weston et al., Memory Networks is used as **static storage** of information

top K

$$o_0 = \xrightarrow{s_o([x], m_i])}$$

Memory Content
m_1
$\boxed{m_2}$
$\ldots$
m_N

Detail

1. Weston et al., Memory Networks, InProc ICLR 2015
2. s_o is a scoring function

P2
Memory Network

MemNN Response

- Decoding module O
  - Selects k supporting memory cells $m_{o_1}, \ldots, m_{o_k}$
    - $o_k = O(x, \mathbf{M}) = \underset{i=0,\ldots,n-1}{\text{argmax}}\ s_o([x, o_{<k}], m_i])$[2]
  - Can also use Attention Mechanism instead
    - $o = \sum_i w_i m_i$

top K

$s_o([x, o_0], m_i])$

$o_0 = m_2$

$o_1 =$

Memory Content
m_1
m_2
$\ldots$
m_i

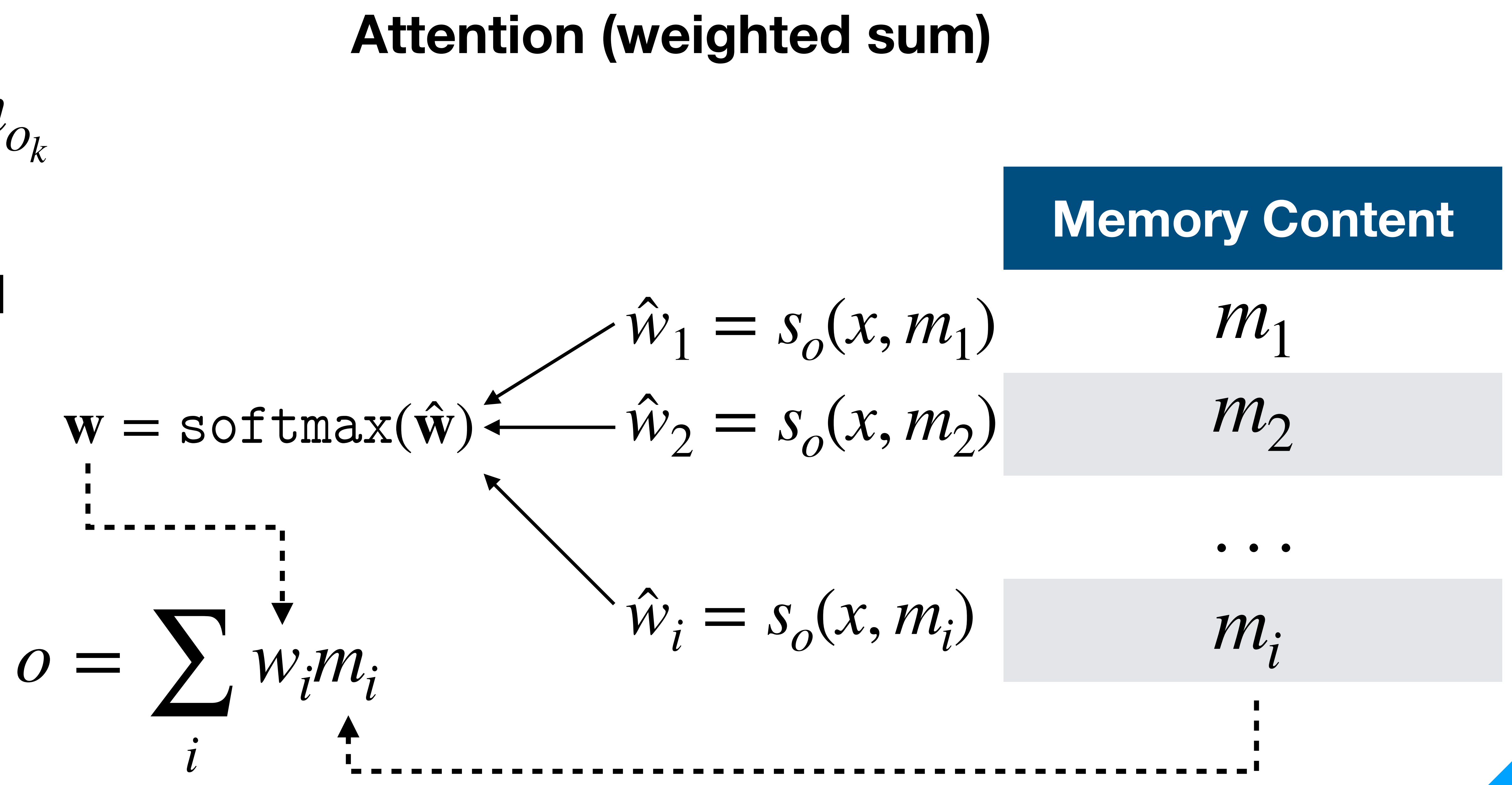
- In Weston et al., Memory Networks is used as **static storage** of information

1. Weston et al., Memory Networks, InProc ICLR 2015
2. s_o is a scoring function

Detail

MemNN Response

- Decoding module O
  - Selects k supporting memory cells $m_{o_1}, \ldots, m_{o_k}$
    - $o_k = O(x, \mathbf{M}) = \underset{i=0,\ldots n-1}{\text{argmax}}\ s_o([x, o_{<k}], m_i])$[2]
  - Can also use Attention Mechanism instead
    - $o = \sum_i w_i m_i$



- In Weston et al., Memory Networks is used as **static storage** of information

1. Weston et al., Memory Networks, InProc ICLR 2015
2. s_o is a scoring function

Detail

P2
Memory Network

Attentional Retrieval

$$\hat{w}_i = s_o(x, m_i)$$

- Say, $s_o(x, y) = x \cdot y$
  - x: query, with sentence representation
    What does NLP stand for?
  - y: memory, with representation for a single memory unit
    NLP stands for Naughty Lousy Parents.
  - $x \cdot y$: information shared between x and y, aside from the similar dimensions
    for x, there's query
    for y, you find features for the response

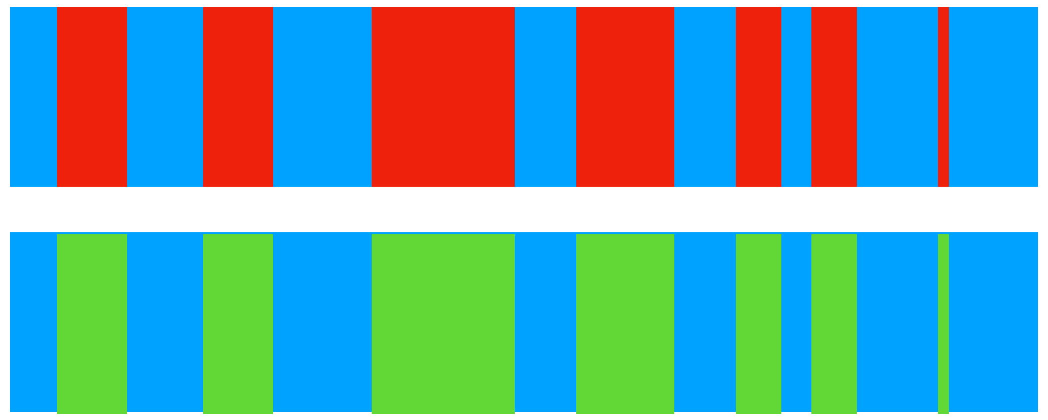
1. A rough example

Detail

P2
Memory Network

Attentional Retrieval

$$\hat{w}_i = s_o(x, m_i)$$



- Say, $s_o(x, y) = x \cdot y$
    - x: query, with sentence representation
      What does NLP stand for?
    - y: memory, with representation for a single memory unit
      NLP stands for Naughty Lousy Parents.
    - $x \cdot y$: information shared between x and y, aside from the similar dimensions
      for x, there's query
      for y, you find features for the response

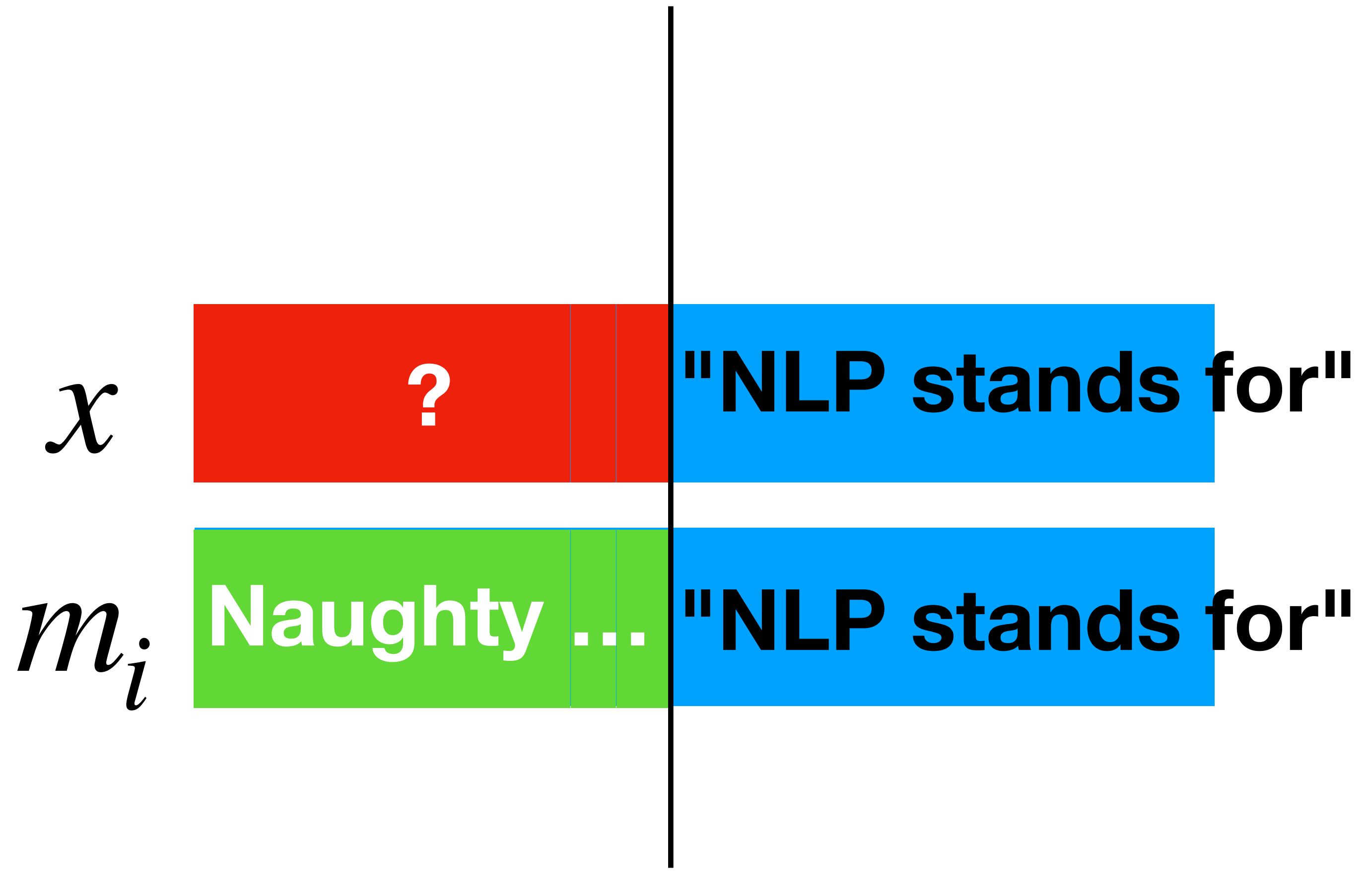
1. A rough example

Detail

P2
Memory Network

Attentional Retrieval

$$\hat{w}_i = s_o(x, m_i)$$



- Say, $s_o(x, y) = x \cdot y$
  - x: query, with sentence representation
    What does NLP stand for?
  - y: memory, with representation for a single memory unit
    NLP stands for Naughty Lousy Parents.
  - $x \cdot y$: information shared between x and y, aside from the similar dimensions
    for x, there's query
    for y, you find features for the response

Detail

1. A rough example

P2
Memory Network

Attentional Retrieval

- Different Attention?
  - Sure, why not
  - E.g. Luong et al. attention
    - $s(x, y) = x^T y$
    - $s(x, y) = x^T W y = x^T (W y)$, projecting y before using above
    - $s(x, y) = V \tanh(W[x; y]) = V[\tanh(W_1 x + W_2 y)]$
      More projections and added activation/normalisation in between

1. A rough example

Detail

Memory Network
MemNN

- MemNN Features
  - First real-world application, trained end-to-end
  - **Efficient Context Processing**
    by new storage format
  - **Massive Storage Capacity**
    A database with 14M facts were used in experiment
  - **Excellent Performance** in Retrieval
    Parallel execution possible

Technical

P2
Memory Network

End-to-End Memory Network (MemN2N)

- Key-Value Memory Network
  - *key* vectors for addressing,
    value vectors for aggregation
- Attentional Read
  - Content-based attentional weight calculation using $k_{1...N}$
    $$w_i = \text{softmax}(s(k, q))_i$$
  - Final read given query q
    $$\sum_i w_i m_i$$

Memory Bank
(k_1, m_1)
(k_2, m_2)
$\dots$
(k_N, m_N)

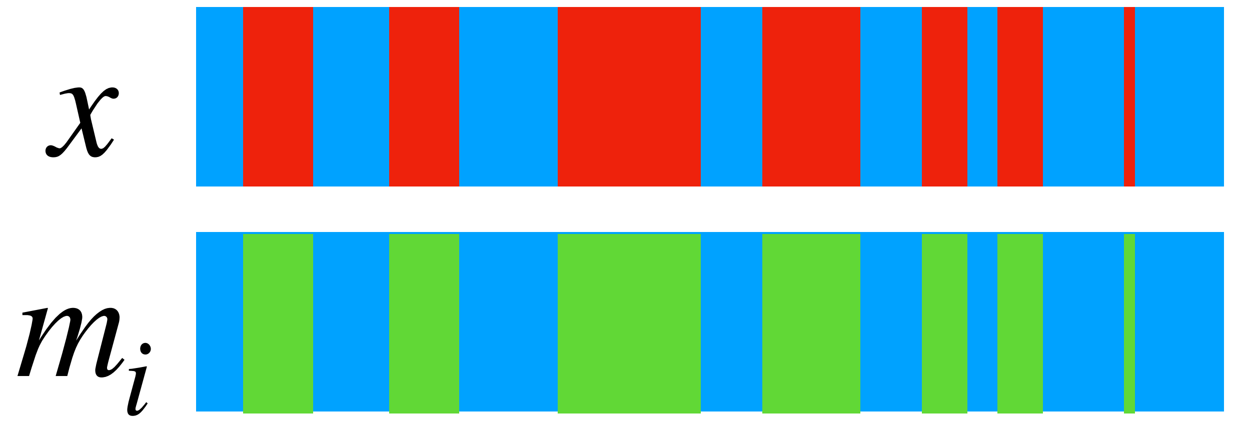
Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

P2
Memory Network

End-to-End Memory Network (MemN2N)

- Key-Value Memory Network
  - *key* vectors for addressing, *value* vectors for aggregation



- Attentional Read
  - Content-based attentional weight calculation using $k_{1 \dots N}$

    $$w_i = \text{softmax}(s(k, q))_i$$

  - Final read given query q

    $$\sum_i w_i m_i$$

Memory Bank
(k_1, m_1)
(k_2, m_2)
$\dots$
(k_N, m_N)

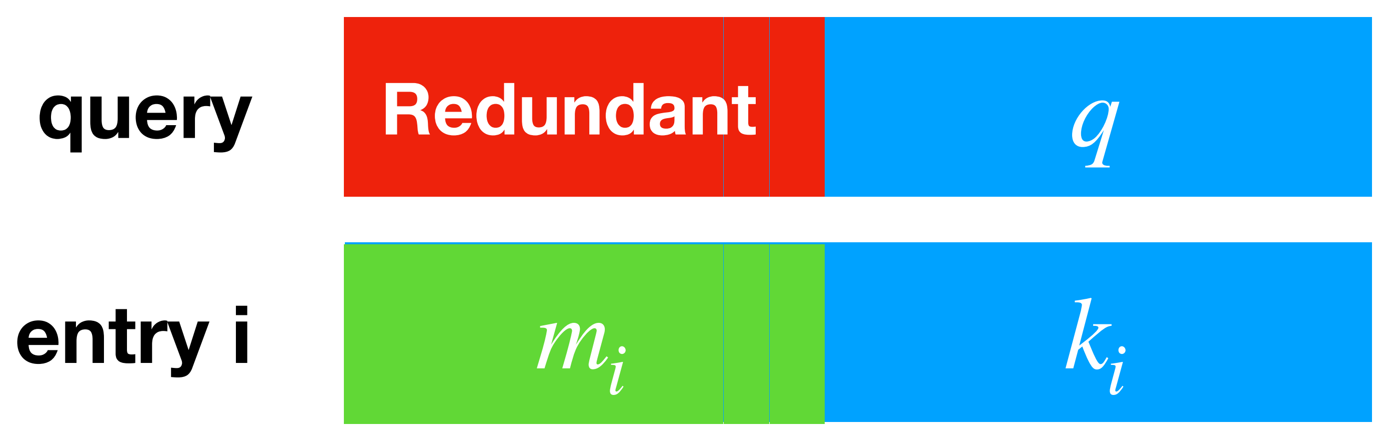
Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

P2
Memory Network

End-to-End Memory Network (MemN2N)

- Key-Value Memory Network
  - *key* vectors for addressing, *value* vectors for aggregation



- Attentional Read
  - Content-based attentional weight calculation using $k_{1...N}$
    $$w_i = \text{softmax}(s(k, q))_i$$
  - Final read given query q
    $$\sum_i w_i m_i$$

Memory Bank
(k_1, m_1)
(k_2, m_2)
$\dots$
(k_N, m_N)

Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

P2
Memory Network

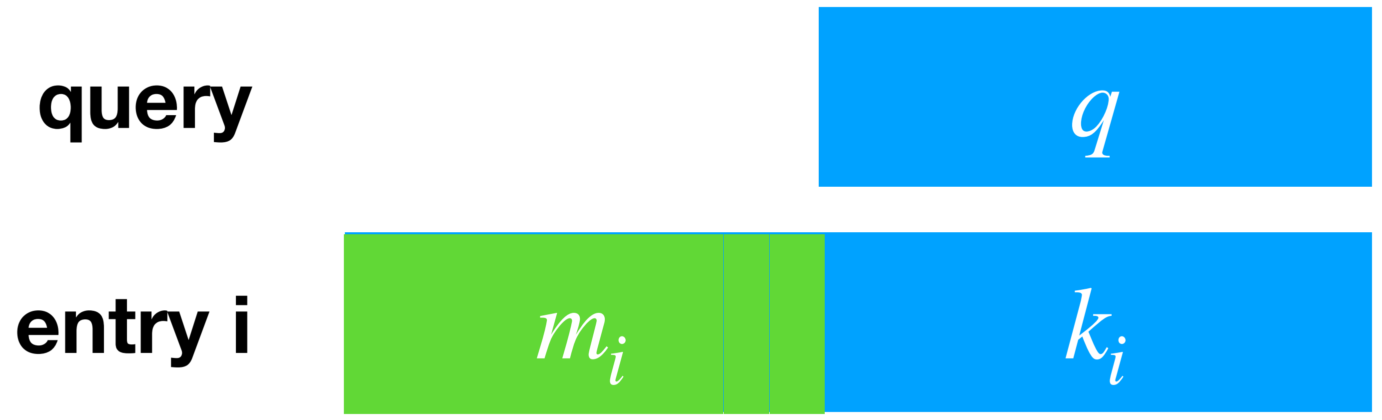
End-to-End Memory Network (MemN2N)

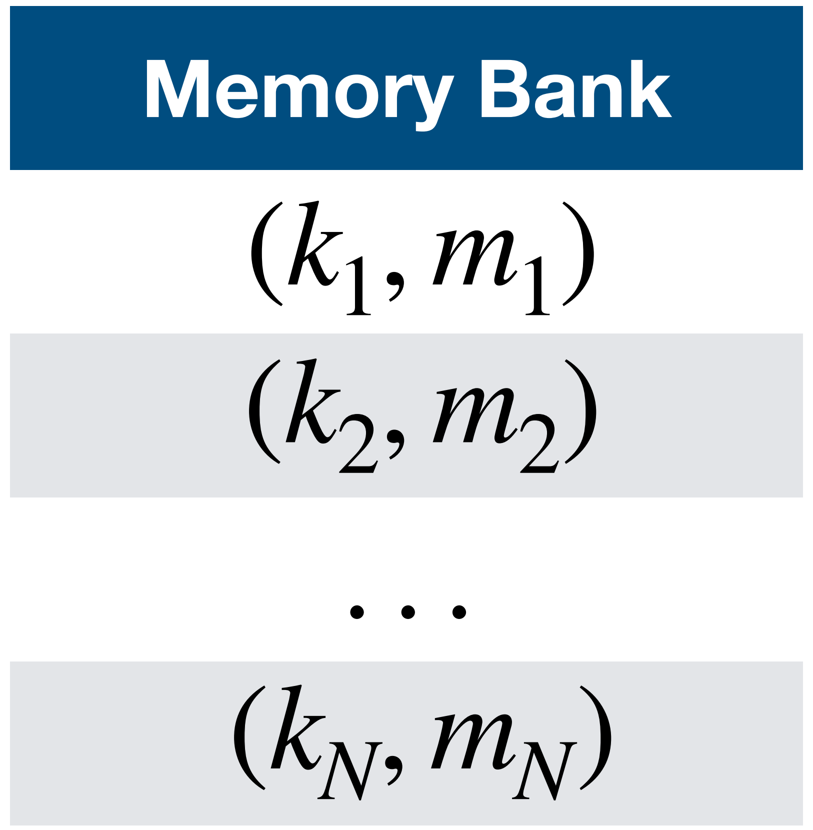
- Key-Value Memory Network
  - *key* vectors for addressing, *value* vectors for aggregation
- Attentional Read
  - Content-based attentional weight calculation using $k_{1...N}$

$$w_i = \text{softmax}(s(k, q))_i$$

  - Final read given query q

$$\sum_i w_i m_i$$





Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

End-to-End Memory Network (MemN2N)

query q

entry i m_i k_i

- Advantages
  - Separation for query information, and actual memory for content
  - Update to key and value separately
  - Easier to train

Memory Bank

(k_1, m_1)

(k_2, m_2)

$\dots$

(k_N, m_N)

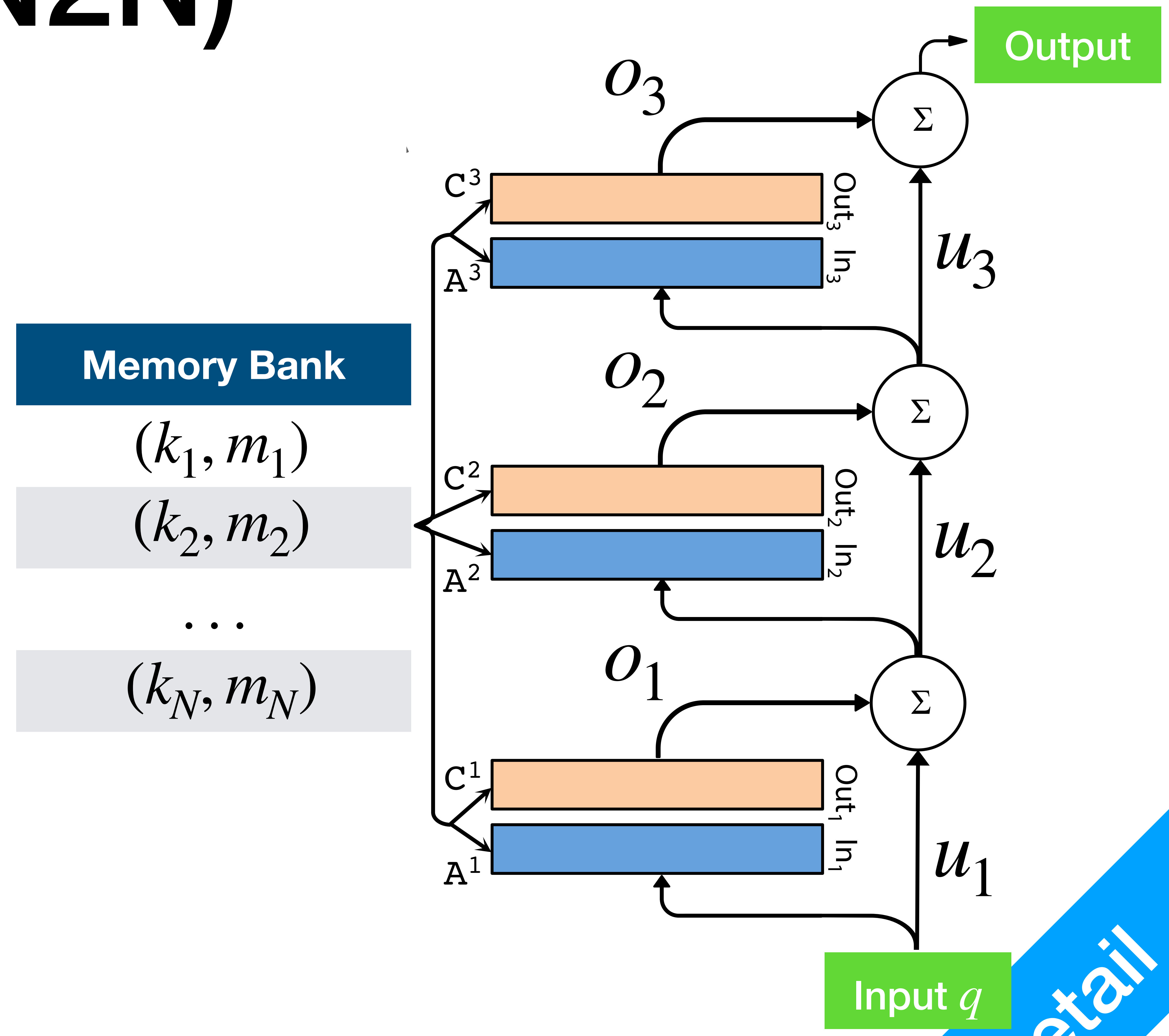
Detail

1. Weston et al., Memory Networks, InProc ICLR 2015

P2
Memory Network

End-to-End Memory Network (MemN2N)

- Decoder goes through multiple passes of retrieval
  - Multi-hop QA, each time different information could be accessed
  - Input at layer k is the combined representation of output o_{k-1} and previous input u_{k-1}
  - u_1 is q encoded



Detail

1. Sukhbaatar et al., End-To-End Memory Networks, InProc NIPS2015

End-to-End Memory Network (MemN2N)

- Decoder goes through multiple layers of retrieval
  - Multi-hop QA, each time different information could be accessed
  - Input at layer k is the combined representation of output o_{k-1} and previous input u_{k-1}
  - u_1 is q encoded

Paragraph

Sandra dropped the milk.

HOP1 John took the milk there.

Sandra went to the bathroom.

HOP2 John moved to the hallway.

Mary went back to the bedroom.

Query

Where is the milk?

Memory Bank
(k_1, m_1)
(k_2, m_2)
$\dots$
(k_N, m_N)

Example

1. Sukhbaatar et al., End-To-End Memory Networks, InProc NIPS2015

End-to-End Memory Network
MemN2N

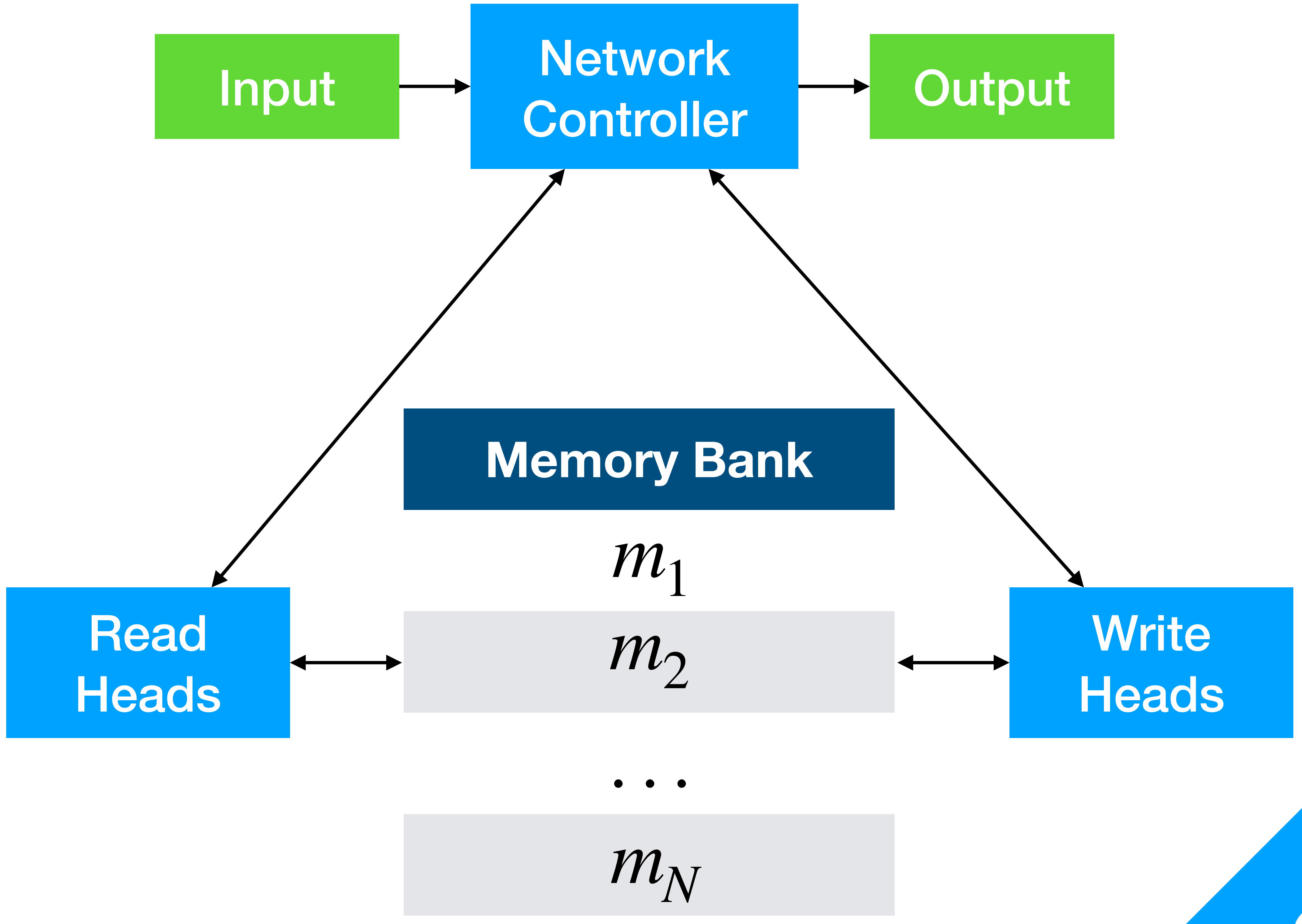
- MemN2N Features
  - Multi-Step Retrieval allows for easier **Multi-Hop Reasoning**
  - Key-Value storage **more practical** than content-based weight calculation using entire memory slots
  - Retrieval tactics:

Technical

P2
Memory Network

Neural Turing Machine

- Storage
  - N slots, of M-dimensional vector each
- Multiple Heads
  - Each Read/Write heads operate **independently** to aggregate information
  - **Multiple Heads are combined** to make up final representations



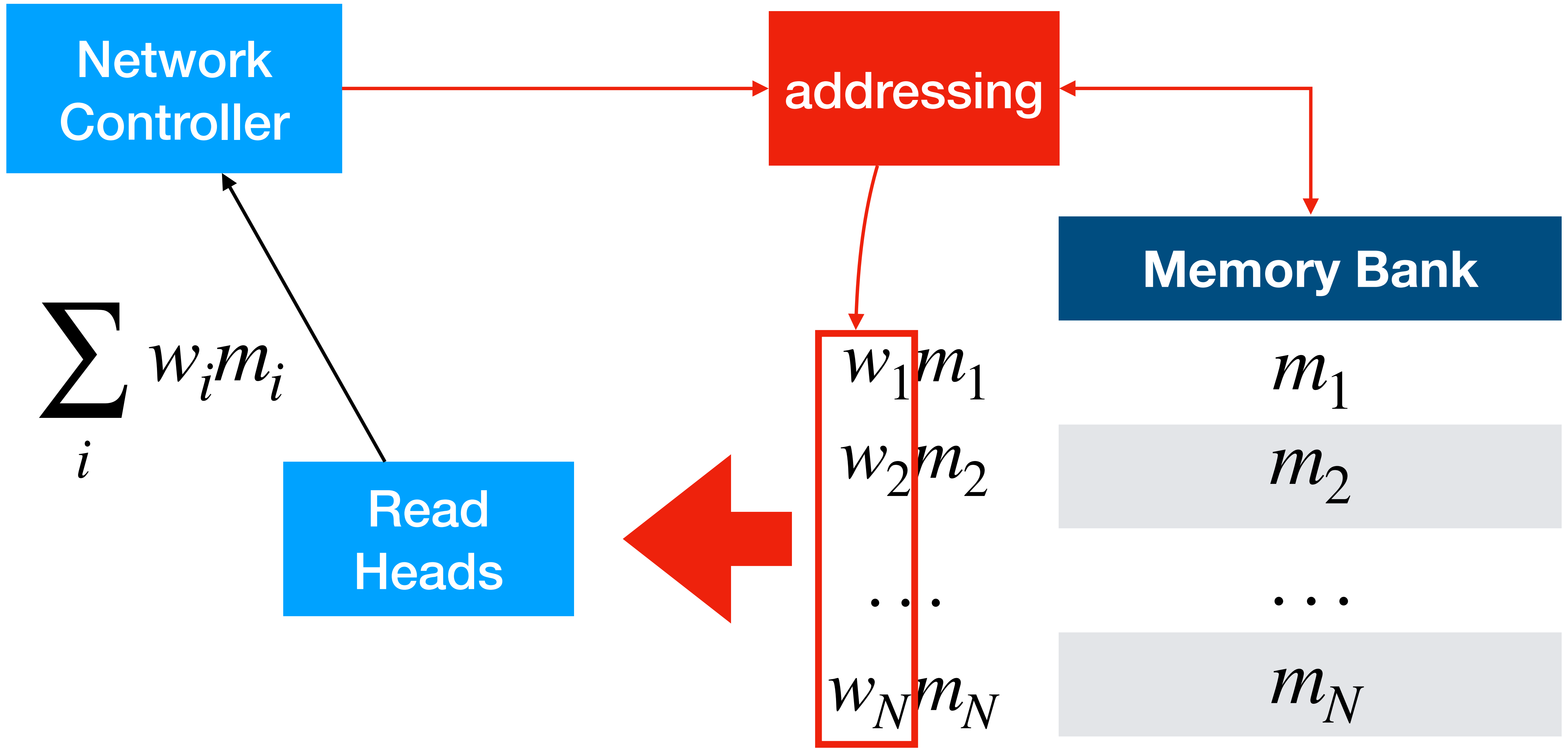
Technical

1. Graves et al. (2014), Neural Turing Machine

P2
Memory Network

Neural Turing Machine

- **Read Operation**
  - Addressing Mechanism provides weights
    - based on Content Cosine Similarity
    - based on Memory location rotational shift of weighting
  - Information aggregated by weighted sum

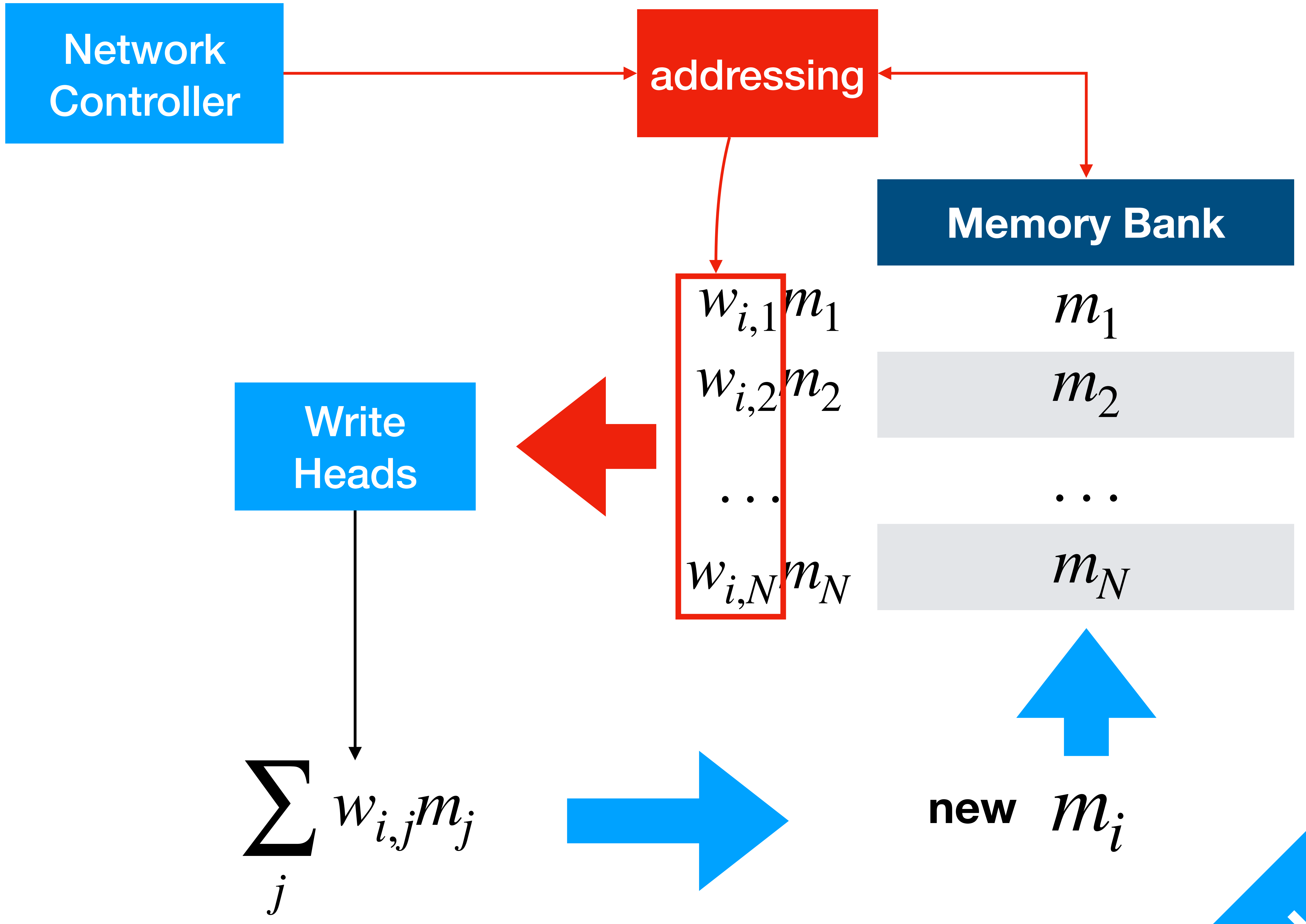


Technical

P2
Memory Network

Neural Turing Machine

- **Write Operation**
  - Aggregation per slot similar to Read
  - At every time step
    - new input q arrives
    - each slot i is updated
    - cost: $O(N^2)$



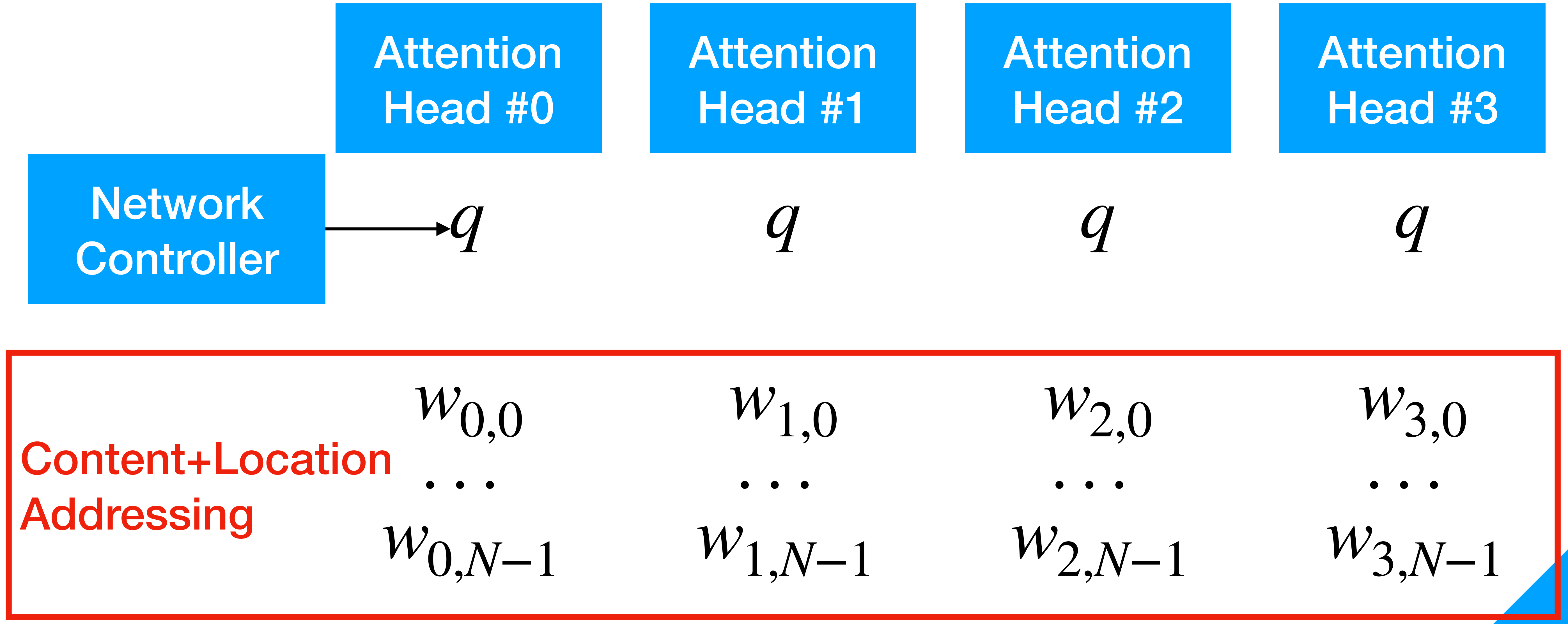
Technical

P2
Memory Network

Neural Turing Machine

- **Example Write Operation**
  - Aggregation per slot similar to Read
  - At every time step
    - new input q arrives
    - each slot i is updated
    - cost: $O(N^2)$

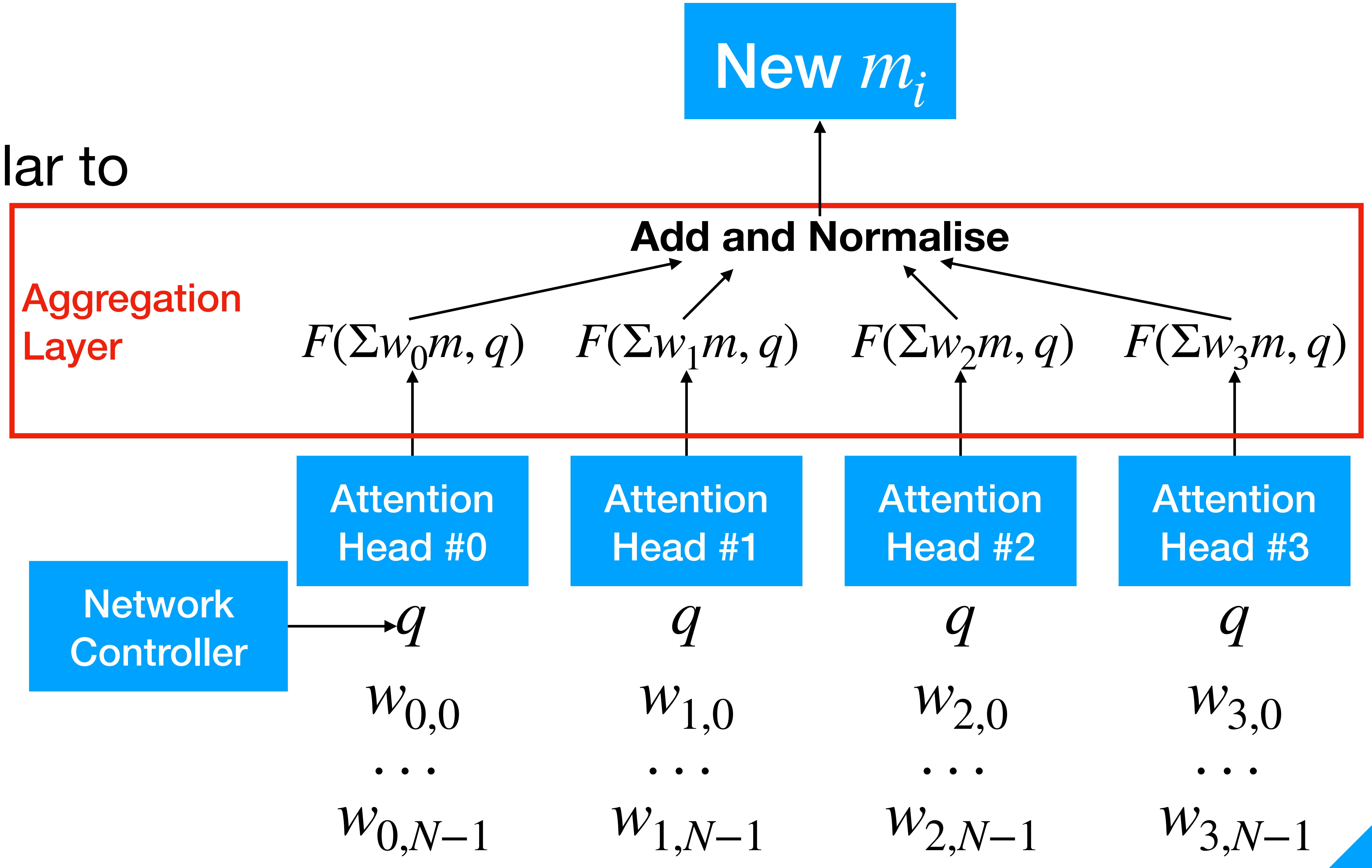
m_i



Technical

Neural Turing Machine

- **Example Write Operation**
  - Aggregation per slot similar to Read
  - At every time step
    - new input q arrives
    - each slot i is updated
    - cost: $O(N^2)$



Technical

P2
Memory Network

Neural Turing Machine

- NTM Features
  - **Distributed memory storage**: each piece of information is stored across entire *memory bank*
  - **Dynamic interaction**: at every time step, each memory slot aggregates information from other slots through Attention
  - **Increased storage capacity**, excellent performance in synthetic tasks

Technical

P2
Memory Network

Memory Networks

	Slot format	One piece of *context*	Cost for adding more *context*	Weight calc.	Information Aggregation	Passes
NTM[1]						
MemNN[2]						
MemN2N[3]						

- Memory Network architecture is **highly modular**
  - Mix and Match components (including Read and Write mechanisms)

1. Graves et al. (2014), Neural Turing Machine
2. Weston et al., Memory Networks, InProc ICLR 2015
3. Sukhbaatar et al., End-To-End Memory Networks, InProc NIPS2015

Technical

Applications of MN

1. **Variety of Context**
   Combination of structured *context* and unstructured textual *context*

2. **Massive Context**
   Integration of massive knowledge base (triplets, graphs, plain-text)

3. **Complex Internal Dynamics**
   Perform complex reasoning tasks

Concept

P3
Improvements

Variety of Context

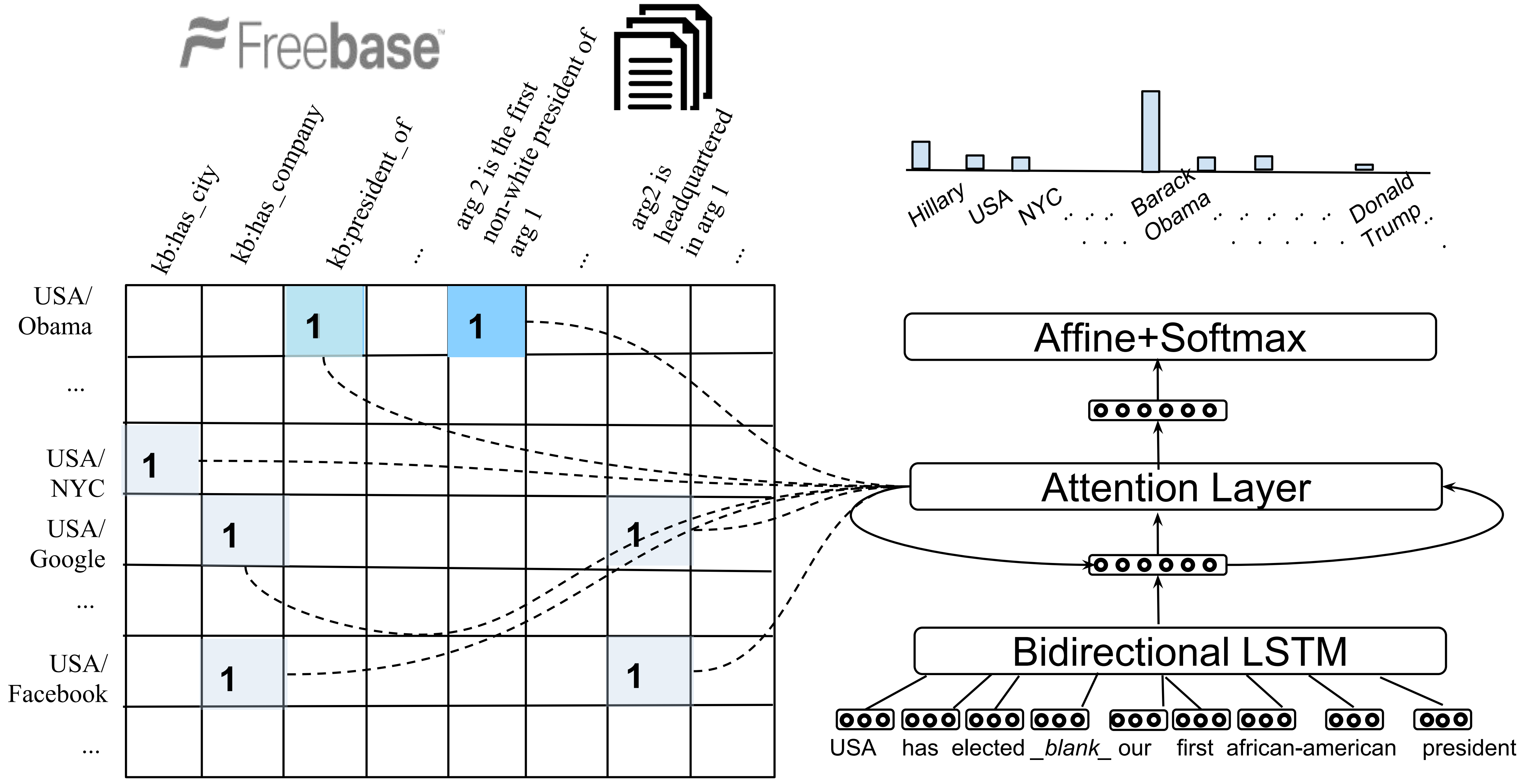


Figure 1: Memory network attending the facts in the universal schema (matrix on the left). The color gradients denote the attention weight on each fact.

- Embed KB facts and text into a uniform representation, as key-value pairs
- Utilise attention mechanism[3] to retrieve information for QA

1. Das et al., Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks, InProc ACL2018
2. Riedel et al., Relation Extraction with Matrix Factorization and Universal Schemas, InProc NAACL-HLT 2013
3. Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate, InProc ICLR2015

Concept

P3
Improvements

Variety of Context

- Universal Schema for KB Triplets (e.g. `(Obama, bornIn, USA)`) and text
  - Sentence have Subject and Object extracted first
    Key: $[E_E(s); \mathbf{LSTM}(Sent)]$, **Value**: $E_E(o)$
  - KB Triplets are embedded with entity and relation concatenated
    Key: $[E_E(s); E_R(r)]$, **Value**: $E_E(o)$

Technical

1. Das et al., Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks, InProc ACL2018
2. Riedel et al., Relation Extraction with Matrix Factorization and Universal Schemas, InProc NAACL-HLT 2013
3. Miller et al., Key-value memory networks for directly reading documents, InProc EMNLP 2017

P3
Improvements

Variety of Context

- Attention mechanism: iteratively generate new context vectors
  - $\mathbf{c}_0$: on the question itself
  - $\mathbf{c}_t$: combine $\mathbf{c}_{t-1}$ with Memory attention
    $\mathbf{c}_t = W_t(\mathbf{c}_{t-1} + W_P \sum_{(k,v \in M)} (c_{t-1} \cdot k)v)$, where W_t contains attention weights

Technical

1. Das et al., Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks, InProc ACL2018
2. Riedel et al., Relation Extraction with Matrix Factorization and Universal Schemas, InProc NAACL-HLT 2013
3. Miller et al., Key-value memory networks for directly reading documents, InProc EMNLP 2017

P3 Improvements

Using Universal Schema[2] in QA

Model	Dev. F_1	Test F_1
Bisk et al. (2016)	32.7	31.4
ONLYKB	39.1	38.5
ONLYTEXT	25.3	26.6
ENSEMBLE.	39.4	38.6
UNISCHEMA	**41.1**	**39.9**

Table 1: QA results on SPADES.

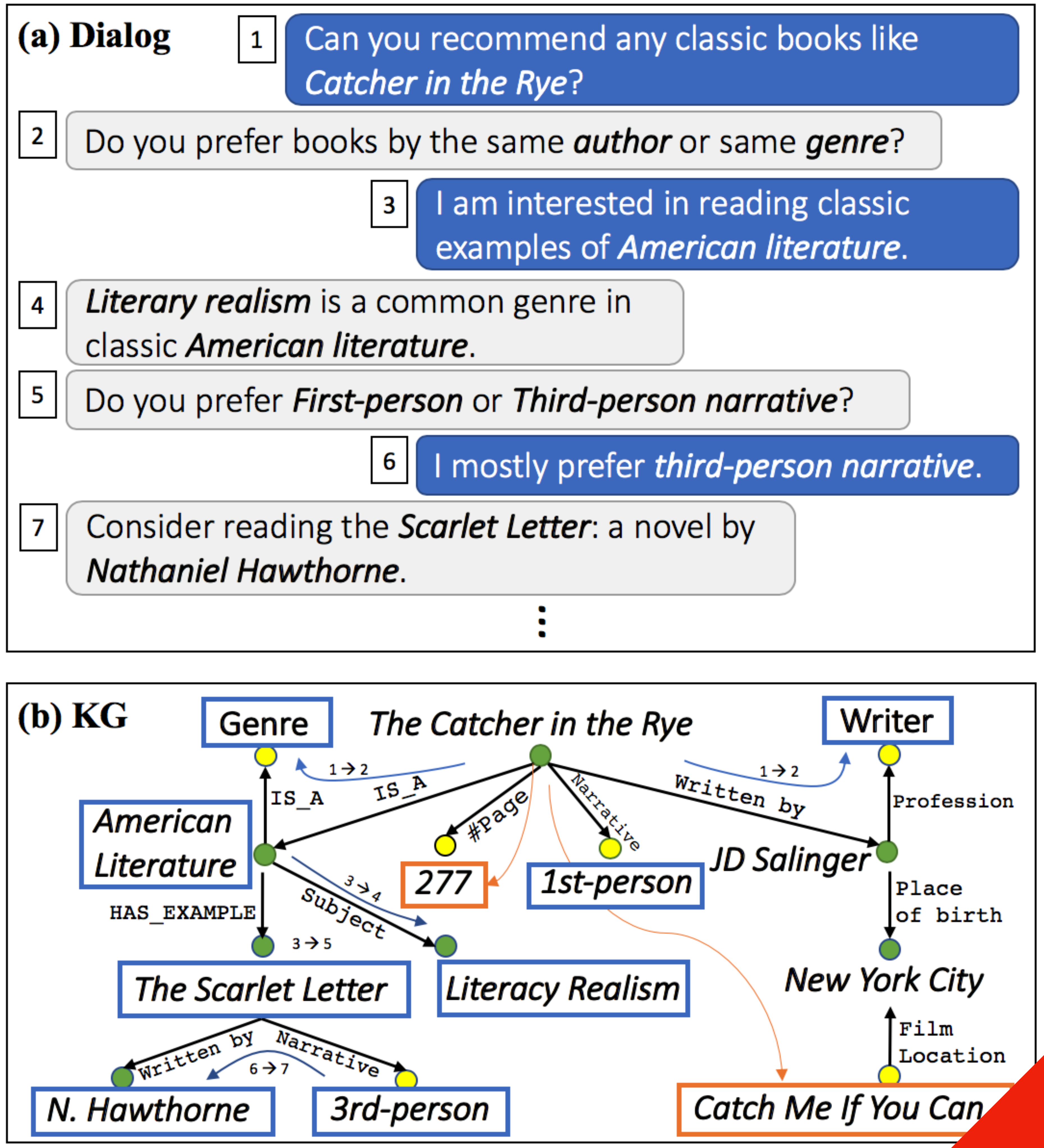
(1) (2)

1. Das et al., Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks, InProc ACL2018
2. Riedel et al., Relation Extraction with Matrix Factorization and Universal Schemas, InProc NAACL-HLT 2013
3. Miller et al., Key-value memory networks for directly reading documents, InProc EMNLP 2017

P3
Improvements

Massive Context

- In reality, one doesn't always get a fine-grained set of factoids for every query
  - Open question answering
  - Open conversation
- Searching for useful information (especially multihop) is difficult in huge knowledge bases
  - Each entity is connected to a lot of other entities, as hops increase the time complexity increases exponentially



Concept

1. Moon et al., OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs., InProc ACL2019

P3
Improvements

Massive Context

- Utilising Random Walk to efficiently retrieve information from a Knowledge Graph
  - A graph could be fully structured (s, r, o) triplet graph
  - A graph could be plain-text connected entity mentions
- Initialisation
  - Utilises TransE to initialise knowledge embedding
  - Knowledge assembled to a graph and encoded to memory cells using Graph Attention
  - Sentence and Dialogue Representation: BiLSTM Encoder and Decoder

1. Bordes et al., Translating Embeddings for Modelling Multi-relational Data., InProc NIPS2013
2. Moon et al., OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs., InProc ACL2019
3. Dhingra et al., Differentiable Reasoning over A Virtual Knowledge Base., InProc ACL2020

Concept

P3
Improvements

Massive Context

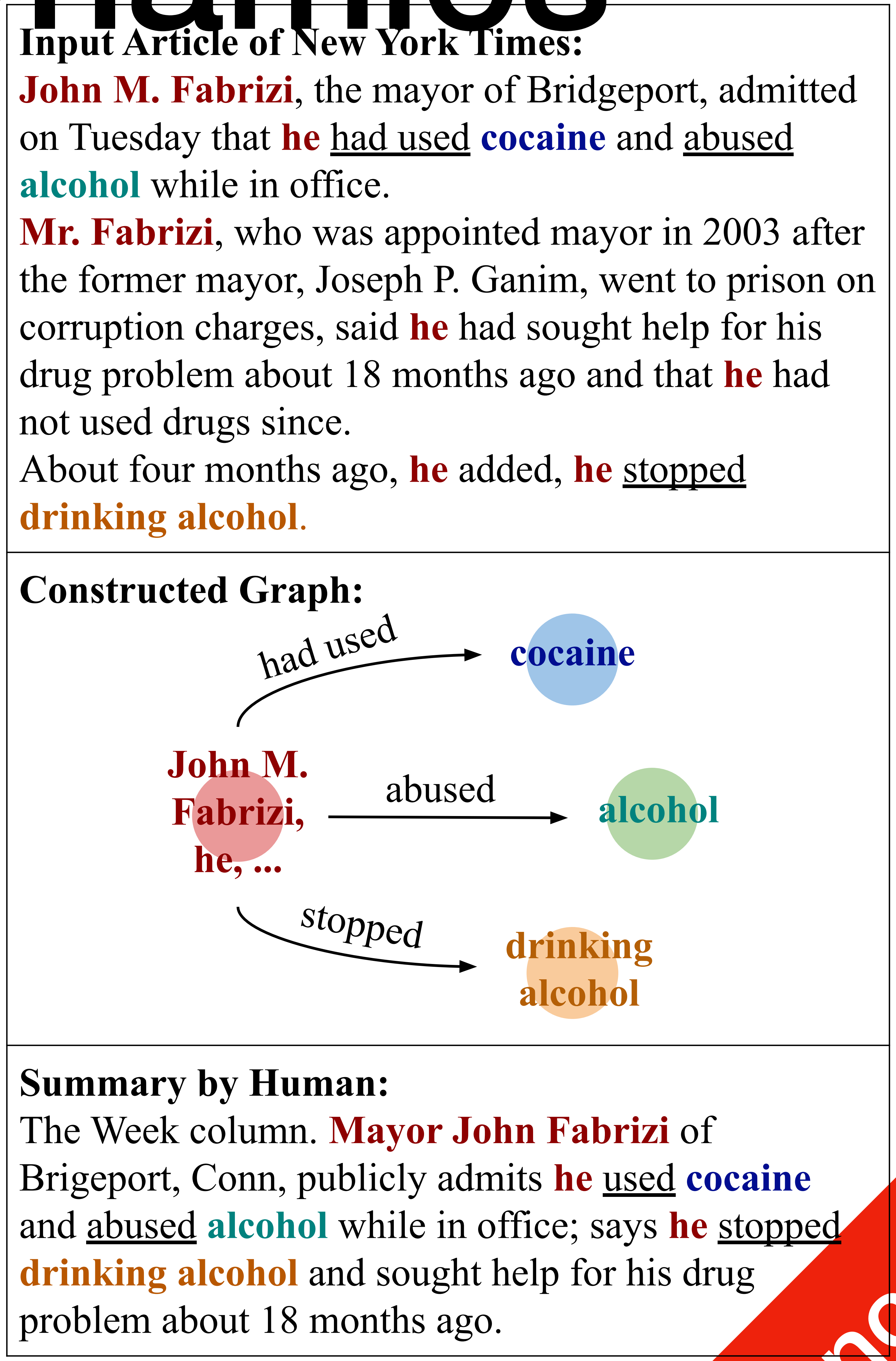
- Random Walk Algorithm
  - Start with an Entity node in a KBG
  - When a query comes in, traverse through connected entities with the highest relevance score
    - In conversation, this help guides the direction of the conversation and retrieve useful information.
    - The path is stored alongside the current context in the decoder LSTM

1. Bordes et al., Translating Embeddings for Modelling Multi-relational Data., InProc NIPS2013
2. Moon et al., OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs., InProc ACL2019

Technical

P3 Improvements

Complex Internal Dynamics



- Treat textual passages as Knowledge Base
  - perform IE to generate a small Knowledge Graph for each Query

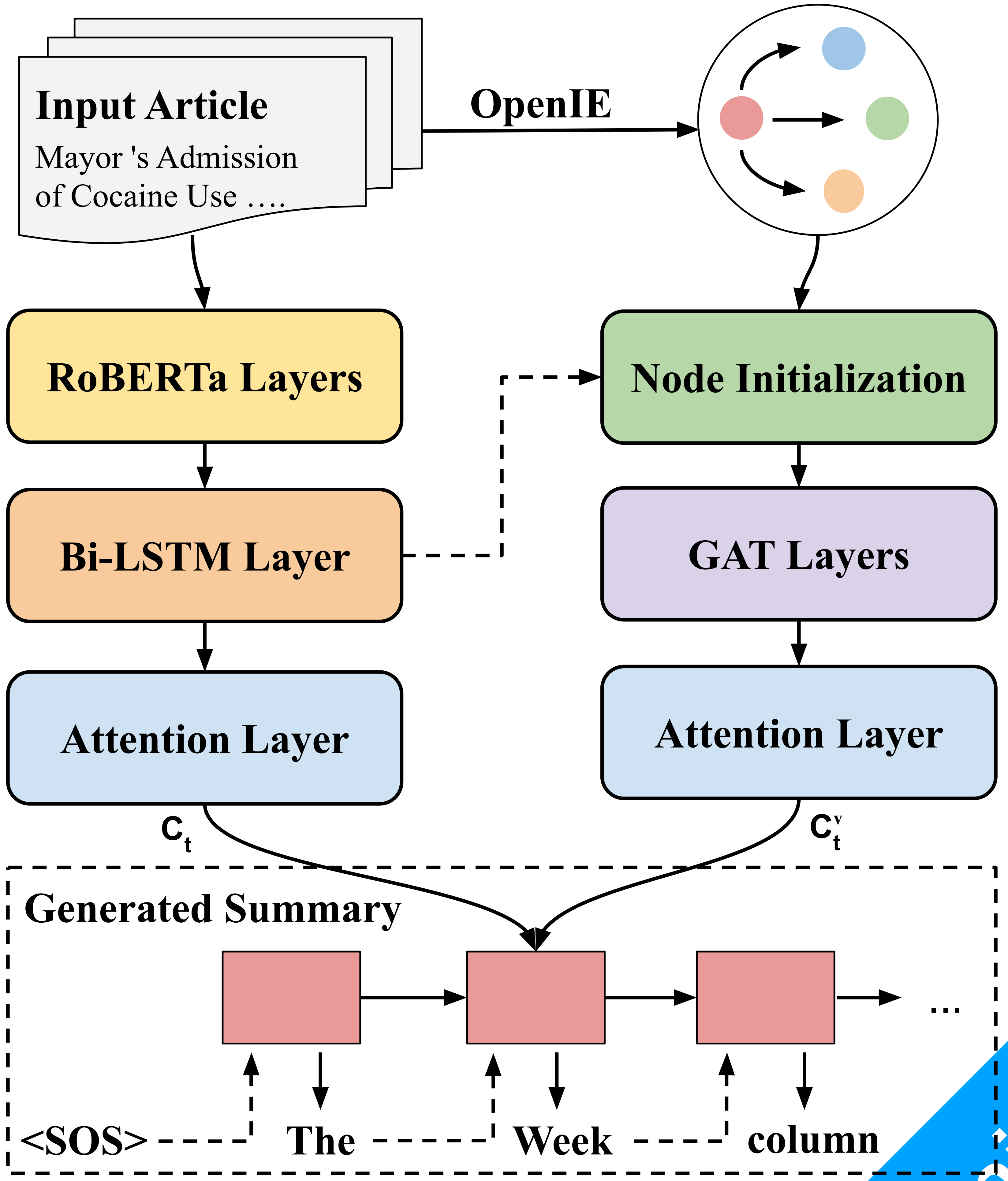
Concept

1. Huang et al., Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward, InProc ACL2020
2. Velicˇkovic et al., Graph Attention Network., InProc ICLR2018

P3 Improvements

Complex Internal Dynamics

- Treat textual passages as Knowledge Base
  - Use BERT to encode input text, and use GAT[2] to encode KB graph as memory cells
  - Use attention to guide summary generation using LSTM



Technical

1. Huang et al., Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward, InProc ACL2020
2. Velicˇkovic et al., Graph Attention Network., InProc ICLR2018

P3 Improvements

Massive KB Graph Integration

Input Dialog (start entity)	Response		
	Model	Walk Path	Predicted Entity
A: *Yes, I believe he [Muller] has played in Munich.* B: *He also won a Bravo Award. I think that's awesome!* A: [response]	GT	award won by → position	Forward
	KG_Walker	award won by	Lionel Messi
	Ext-ED	award won by	Muller
A: *Could you recommend a book by Mark Overstall?* B: [response]	GT	wrote → has genre	Romance
	KG_Walker	wrote → has genre	Romance
	Ext-ED	language	English
A: *Do you like Lauren Oliver. I think her books are great!* B: *I do, Vanishing Girls is one of my favorite books.* A: [response]	GT	written by → wrote	Requiem
	KG_Walker	written by → wrote	Annabel
	Tri-LSTM	released year	2015
A: *What about the Oakland Raiders?* B: *Oh yes, I do like them. I've been a fan since they were runner-up in Super Bowl II. What about you?* // A: [response]	GT	Champion	Packers
	KG_Walker	Champion	Packers
	seq2seq	Runner-up → Is_A	NFL Team
A: *Do you like David Guetta? I enjoy his music.* B: *Oh, I love his lyrics to Love is Gone and the song Wild Ones. What are your favorites?* // A: [response]	GT	composer → composed	Club Can't Handle Me
	KG_Walker	composer → composed	I Love It
	Tri-LSTM	composer	David Guetta

Table 4: **Error analysis**: DialKG Walker with attention (ours) vs. baselines. Ground-truth response (GT) and model predictions of walk paths and future entities for the underlined entity mentions are shown. Dialogs are only partially shown due to space constraints.

1. Moon et al., OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs., InProc ACL2019

Technical

Conclusions

- Advantages of Memory Networks
  - Knowledge **in Neural Space**
    variety of formats/sources
  - Easily **Expandable**
    Storage not limited to small params
  - **Complex Reasoning**
    Including multi-hop logical inferences

Memory Content
m_1
m_2
$\cdots$
m_N

Summary

Future Work

- Advantages of Memory Networks
  - Knowledge **in Neural Space**
    variety of formats/sources
  - Easily **Expandable**
    Storage not limited to small params
  - **Complex Reasoning**
    Including multi-hop logical inferences

- Further research
  - **Distributed Mass Knowledge**
    currently only in NTM
  - **More Efficient Integration**
    Memory update are slow now
  - **More Complex Dynamics**
    Reasoning ability far from human

Future