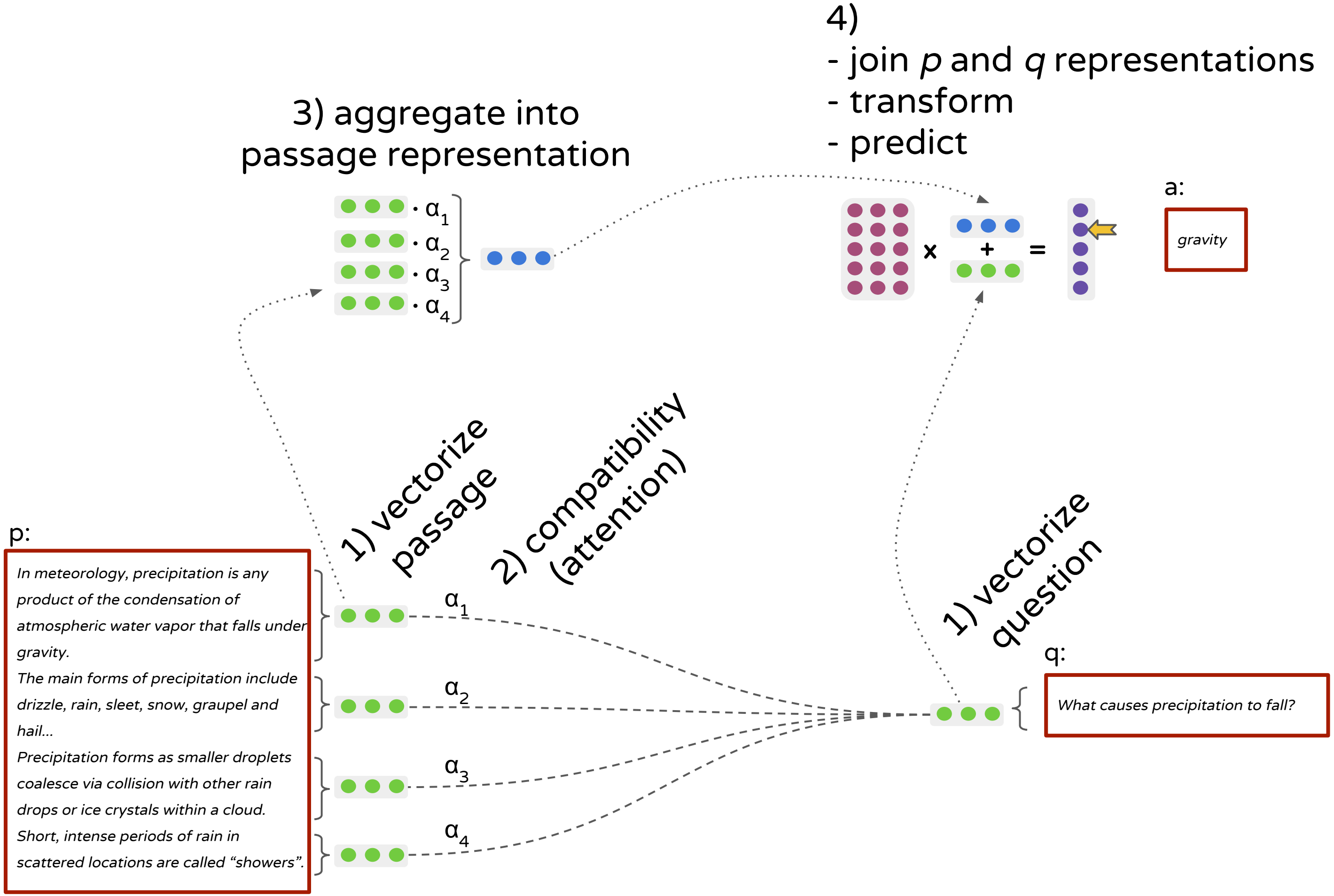
Memories are made of this

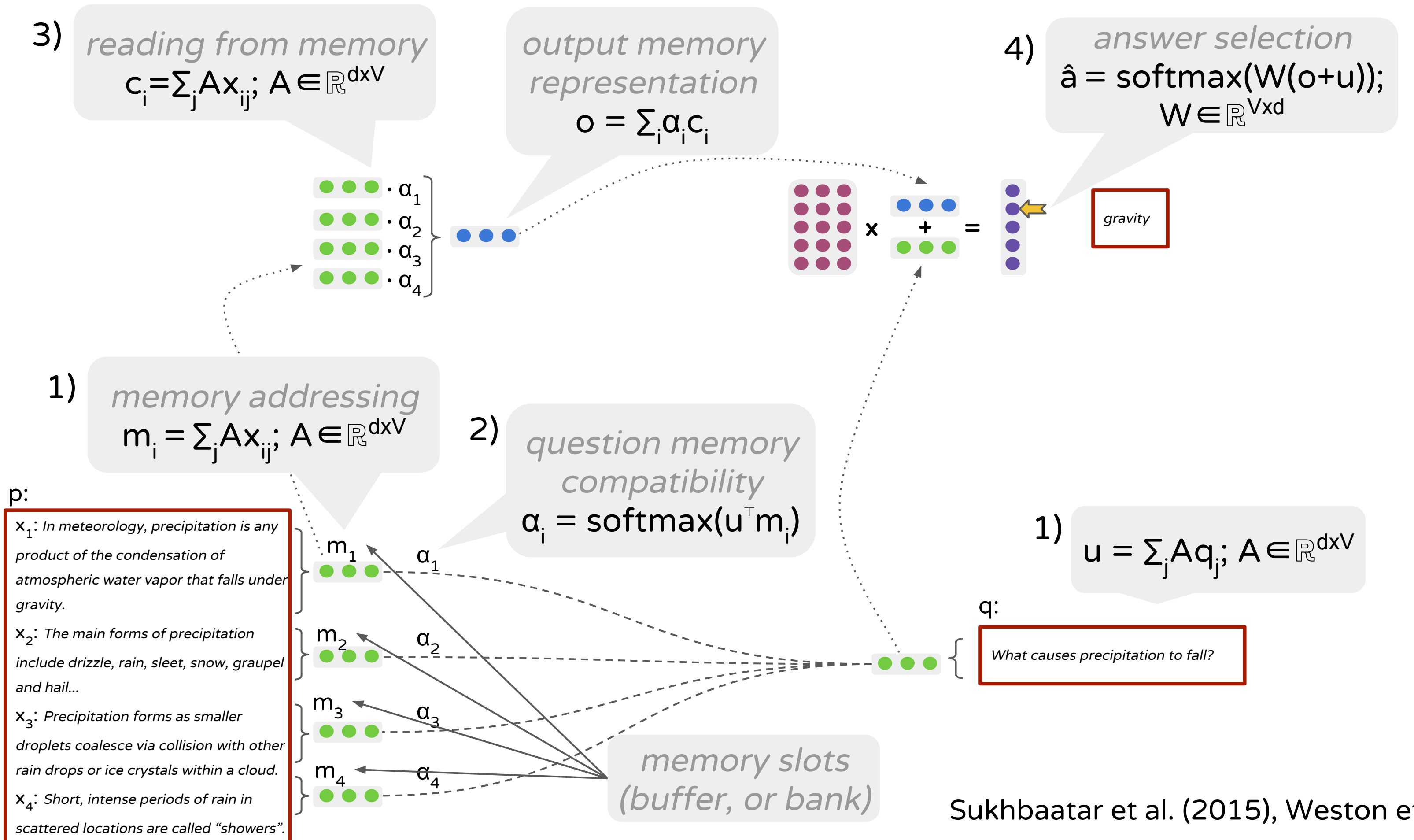
A primer on memory networks for QA

Simon Šuster
CLiPS, 2019

A simple memory network for QA

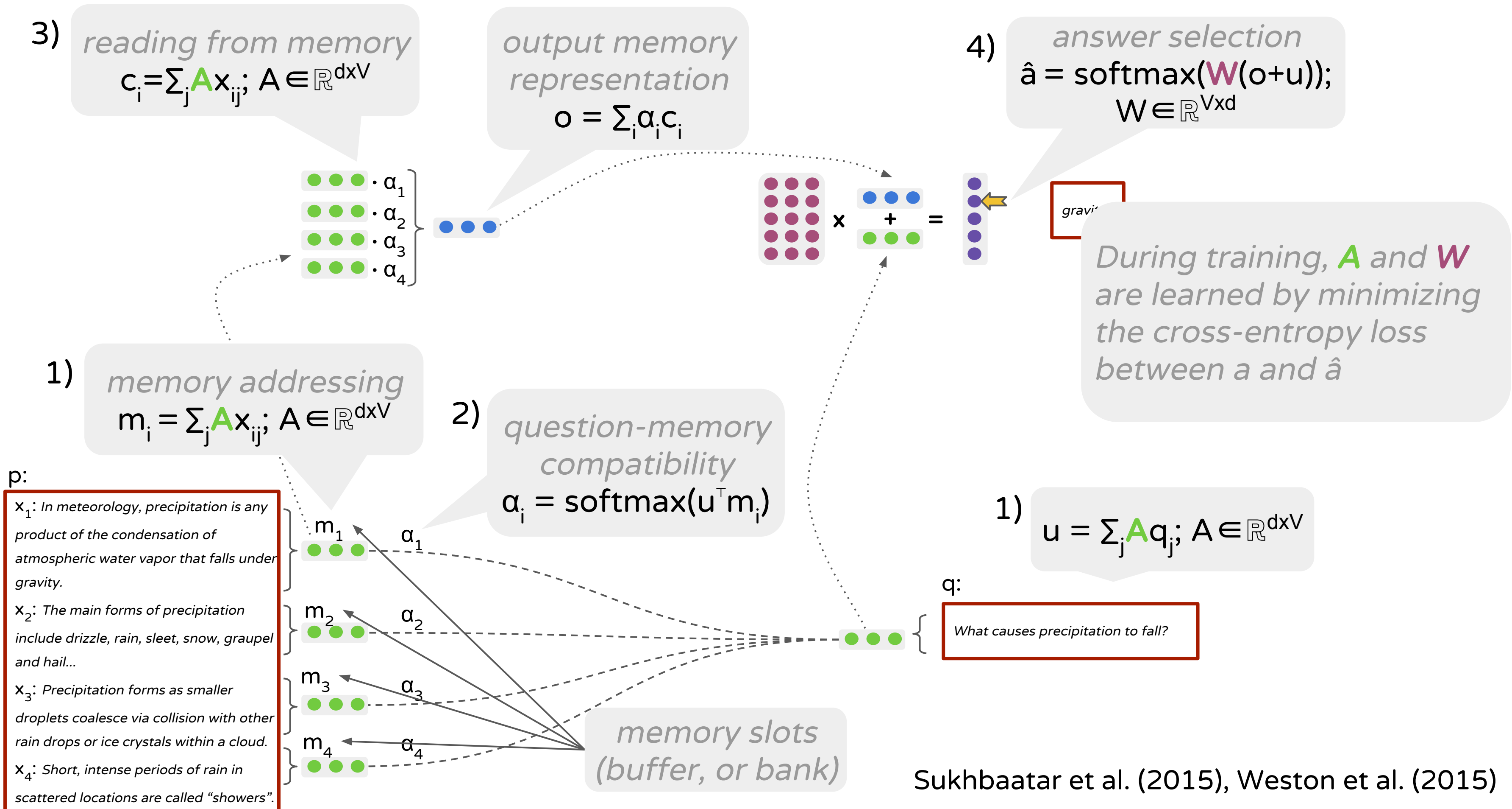


A simple memory network: more details



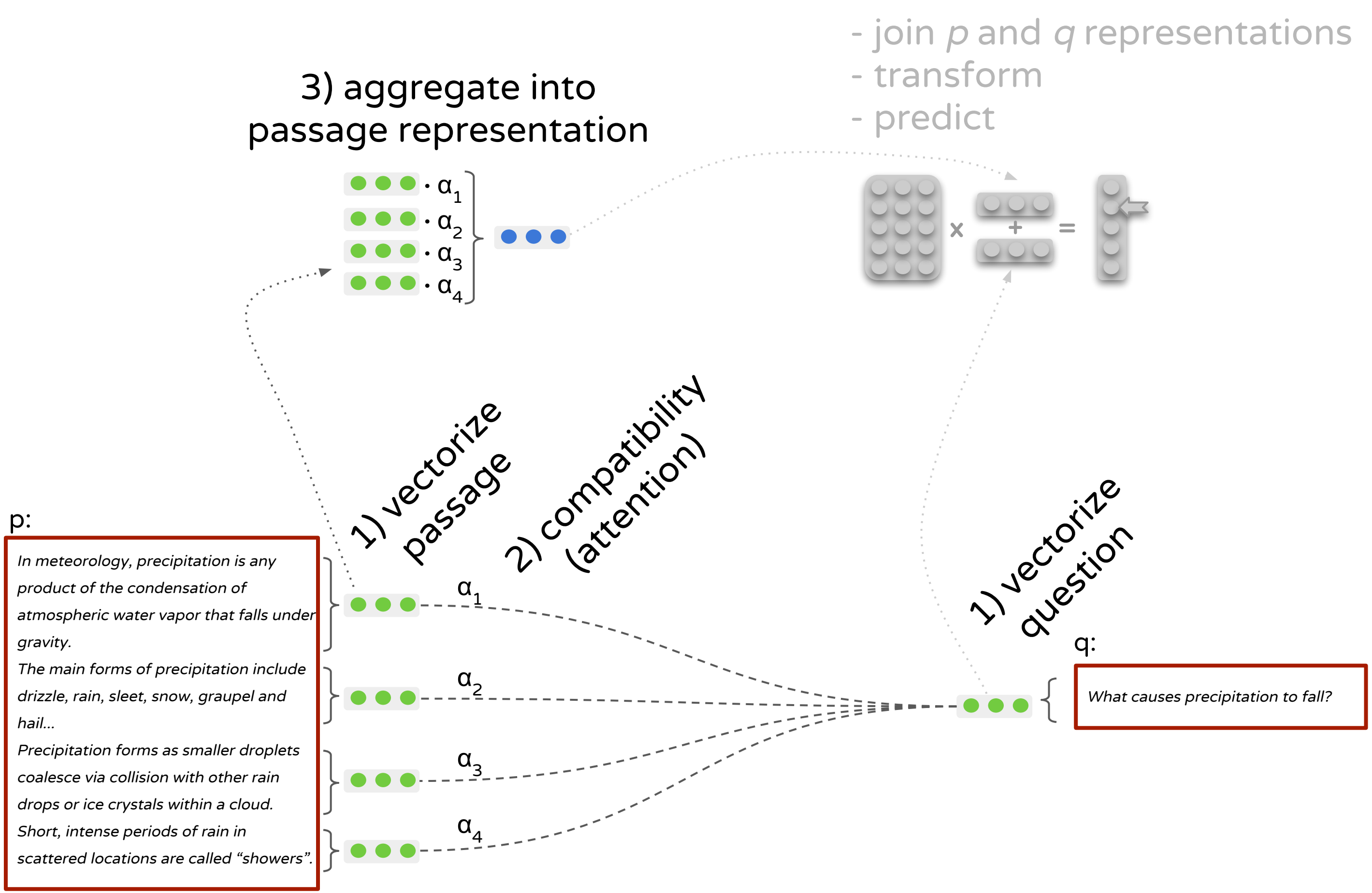
Sukhbaatar et al. (2015), Weston et al. (2015)

A simple memory network: more details



Sukhbaatar et al. (2015), Weston et al. (2015)

A simple memory network for QA: **adding depth**

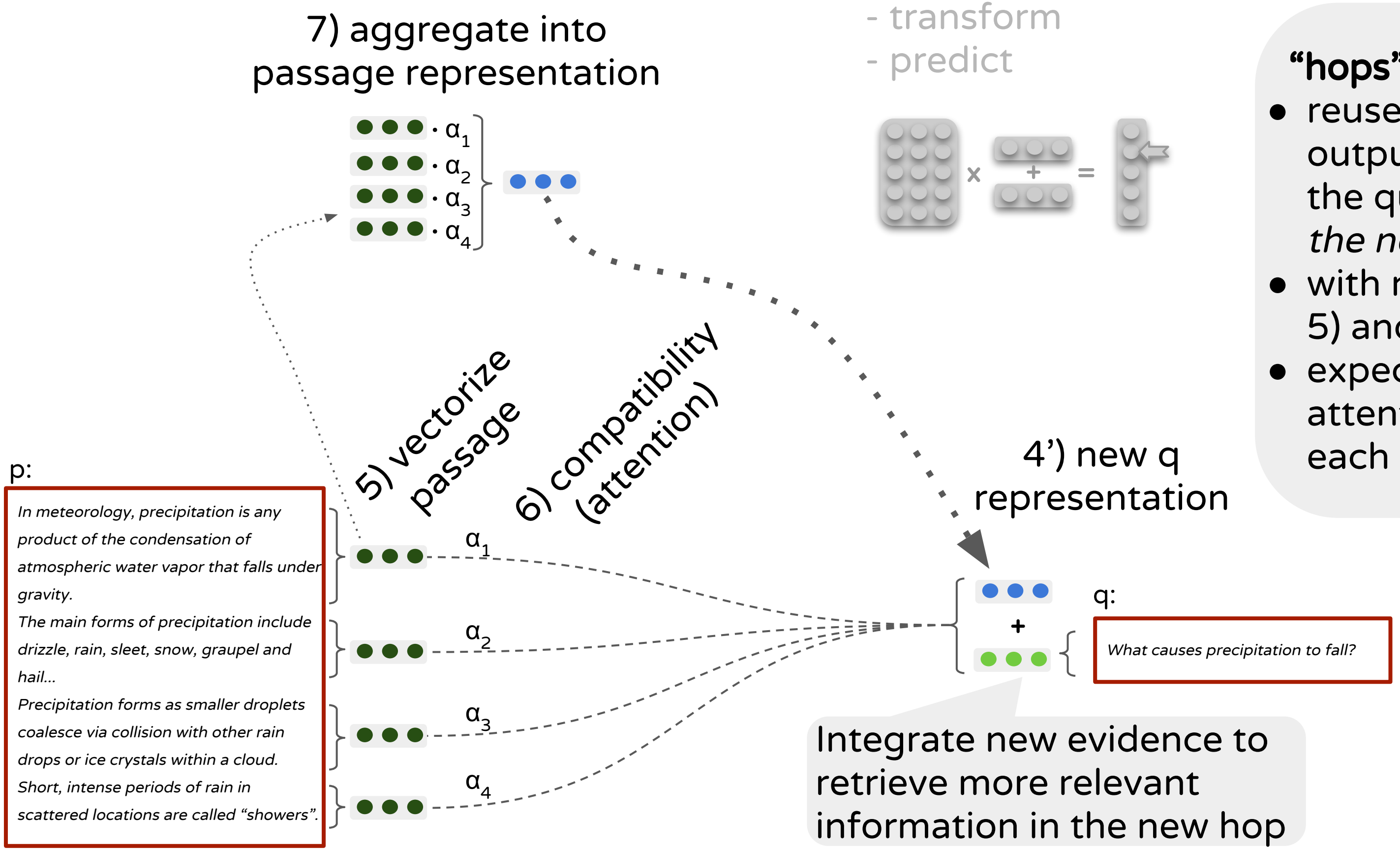


A simple memory network for QA: **adding depth**

- join *p* and *q* representations
- transform
- predict

"hops" (~layers):

- reuse the memory output together with the question vector *in the next pass*
- with new parameters in 5) and 7)
- expect a different attention distribution in each hop



More on multi-step reasoning

Two-step reasoning in bAbI (path-finding):

1 The garden is west of the bathroom.

2 The bedroom is north of the hallway.

3 The office is south of the hallway.

4 The bathroom is north of the bedroom.

5 The kitchen is east of the bedroom.

6 How do you go from the bathroom to the hallway?

s,s   4,2

More on multi-step reasoning

Two-step reasoning in bAbI (path-finding):

1 The garden is west of the bathroom.

2 The bedroom is north of the hallway.

3 The office is south of the hallway.

4 The bathroom is north of the bedroom.

5 The kitchen is east of the bedroom.

6 How do you go from the bathroom to the hallway?

s,s 4,2

*expect strong attention on sent. 2 in the **second hop***

*expect strong attention on sent. 4 in the **first hop***

A high level view: adding depth with multiple hops

Like LSTMs, multihop MemNNs have

- memory and
- recurrency.

But there are some differences wrt to memory and recurrency in both models:

- LSTM memory is **internal** (the state), rewritten at each step with forget/input gates,
- LSTM memory is constantly updated in the activation space, making it potentially more fragile,
- MemNNs give us free hands to define our memory,
- LSTM steps are given by the sequence,
- we can't increase the size of the LSTM memory without increasing the size of the network (computation).

Structuring the memory: ordinary MemNN

a:

gravity

p:

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.
The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail…
Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud.
Short, intense periods of rain in scattered locations are called "showers".

q:

What causes precipitation to fall?

Structuring the memory: windows around entities

values

z_1: *Barack Obama*
z_2: *Illinois*
z_3: *Mitt Romney*
z_4: *Columbia University*

entities

keys

x_1: *____ is the first non-white president of USA.*
x_2: *He previously served as a senator from ____ from 2005 to 2008.*
x_3: *He was subsequently elected to a second term over former Massachusetts governor ____.*
x_4: *After graduating from ____ in 1983, he worked as a community organizer in Chicago.*

a text window around the entity

a:

Barack Obama

q:

America has elected ____, our first African-American president

Structuring the memory: KB triples

values

z_1: *Ridley Scott*
z_2: *Philip K. Dick...*
z_3: *Harrison Ford, ...*
z_4: *1982*

KB objects

a:

Ridley Scott

keys

x_1: *Blade Runner directed_by ____*

x_2: *Blade Runner written_by ____*

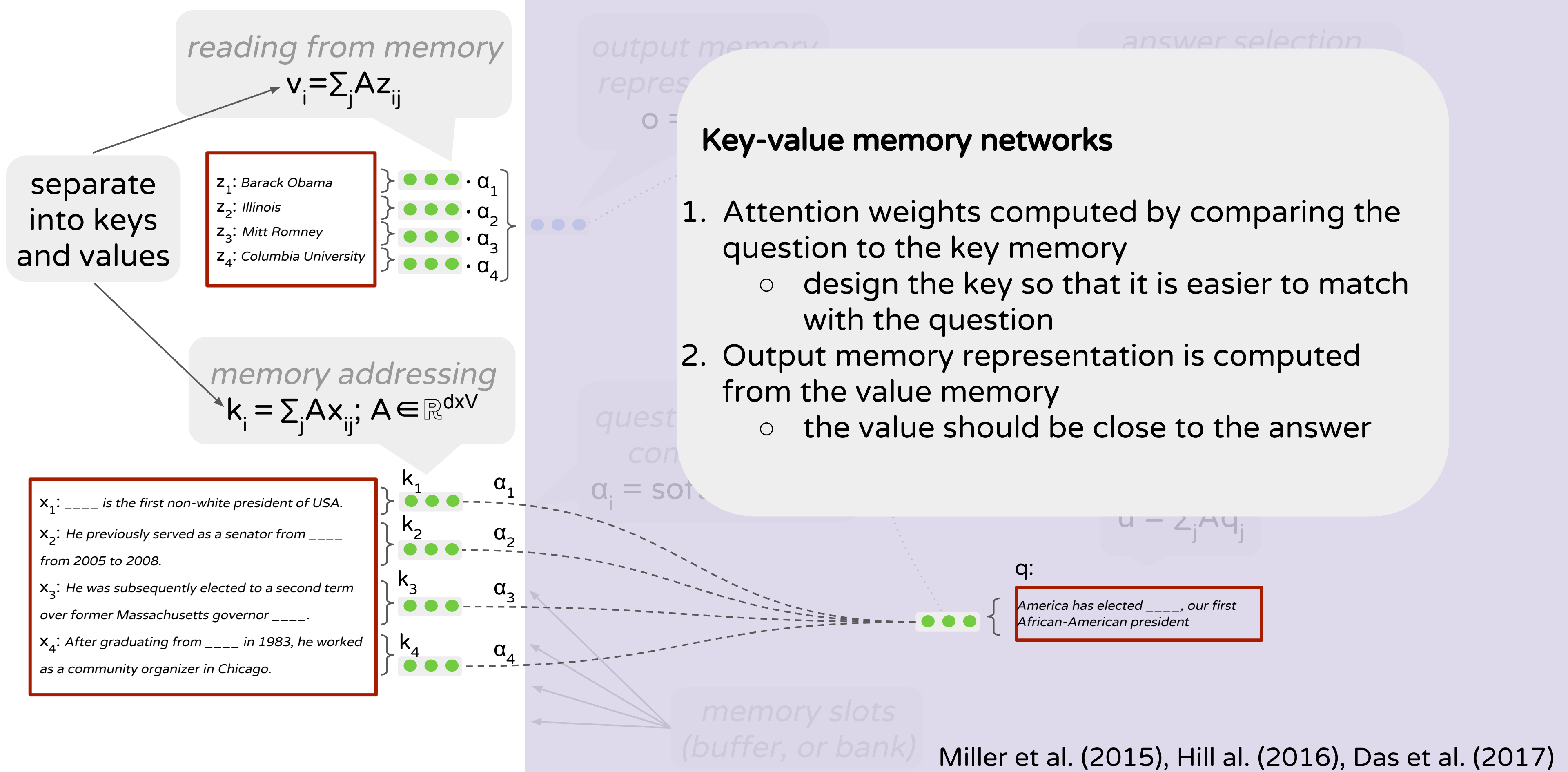
x_3: *Blade Runner starred_actors ____*

x_4: *Blade Runner release_year ____*

KB subjects + relations

Who is the director of the film Blade Runner?

Structuring the memory



Key-value memory networks

1. Attention weights computed by comparing the question to the key memory
   - design the key so that it is easier to match with the question
2. Output memory representation is computed from the value memory
   - the value should be close to the answer

Miller et al. (2015), Hill al. (2016), Das et al. (2017)

Details to determine

- How to embed sentences/input text
  - BoW (+position encoding)
  - LSTMs...
- Share parameters or not
  - Question encoding, memory addressing and memory reading can use distinct parameters
- Parametrize attention?
- Shape of the output layer
  - multiclass over answer vocabulary
  - multiclass over document positions (pointer)
  - RNN
- How to organize the memory
  - flexible!
- How to fit things in the memory (hashing)

Most common use cases

Primarily **QA** datasets: *bAbI*, *WikiHop*, *CNN*, *Children's book test, ...*

Also **visual QA**: *CLEVR*

- decomposing a question into operations that retrieve information from the image (KB) and adding it to the memory state (Hudson and Manning, 2018)

Dialogue: *bAbI*, *Stanford multi-domain dialogue (SMD)*, *Dialog State Tracking Challenge (DSTC2)*

- MemNN as an encoder of the dialogue history, coupled with a decoder that reads and copies the memories to generate a response (Madotto et al. 2018)

Limitations

Based on own experiments, cf. also Chen and Durrett (2019):

- model often attends to the wrong memories
- end-to-end training difficult, need supervision at the level of attention during training
- good performance on a dataset doesn't mean the model *can actually* perform multi-hop reasoning

Temporal dependencies between (attended) memories (Wu et al. 2018)

Some other neural models with external memory

Differentiable neural computer (DNC) (Graves et al. 2016)

- not only have content-based memory addressing but also location-based (allowing modification of content-based memory)
- can not only read from but also write to memory (with the mechanism to read memories that were more recently written to)

Dynamic memory networks (Kumar et al. 2016)

- Hops implemented with a GRU model (creating "episodes")
- Reduce other tasks to QA (e.g. sentiment analysis)

Memory, attention and composition (MAC) cell (Hudson and Manning, 2018)

- Similar to DNC but with a recurrent memory structure

Chen, J., and Durrett, G. How to learn (and how not to learn) multi-hop reasoning with memory networks. Submitted to ICLR, 2019.

Das, R., Zaheer, M., Reddy, S., and McCallum, A. Question answering on knowledge bases and text using universal schema and memory networks. ACL, 2017.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A.P.. Hybrid computing using a neural network with dynamic external memory. Nature. 2016 Oct;538(7626):471.

Hill, F., Bordes, A., Chopra, S., and Weston, J. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. ICLR, 2016.

Hudson, D. A., and Manning, C. D. Compositional attention networks for machine reasoning. ICLR, 2018.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. Ask me anything: Dynamic memory networks for natural language processing. ICML, 2016.

Madotto, A., Wu, C., and Fung, P. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. ACL, 2018.

Miller, A., Fisch, A., Dodge, J., Karimi, A. H., Bordes, A., & Weston, J. Key-value memory networks for directly reading documents. EMNLP, 2016.

Sukhbaatar, S., Weston, J., Fergus, R. End-to-end memory networks. In Advances in neural information processing systems, 2015.

Weston, J., Chopra, S., and Bordes, A. Memory Networks. ICLR, 2015.

Wu, C., Madotto, A., Winata, G. I., and Fung, P. End-to-End Dynamic Query Memory Network for Entity-Value Independent Task-Oriented Dialog. ICASSP, 2018.

https://diegma.github.io/slides/ULL2016/MN_NTM.pdf

Attention shifts per hop (Sukhbaatar et al. 2015)

Story (11: basic coherence)	Support	Hop 1	Hop 2	Hop 3
Mary journeyed to the hallway.		0.00	0.01	0.00
After that she journeyed to the bathroom.		0.00	0.00	0.00
Mary journeyed to the garden.		0.00	0.00	0.00
Then she went to the office.		0.01	0.06	0.00
Sandra journeyed to the garden.	yes	0.97	0.42	0.00
Then she went to the hallway.	yes	0.00	0.50	1.00
Where is Sandra? Answer: hallway Prediction: hallway				