Welcome
2017 Faculty Summit
Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI

Microsoft Research
Faculty Summit **2017**

Machine Reading for Question Answering: from symbolic to neural computation

Jianfeng Gao, Rangan Majumder and Bill Dolan

Microsoft AI & R

7/17/2017

Outline

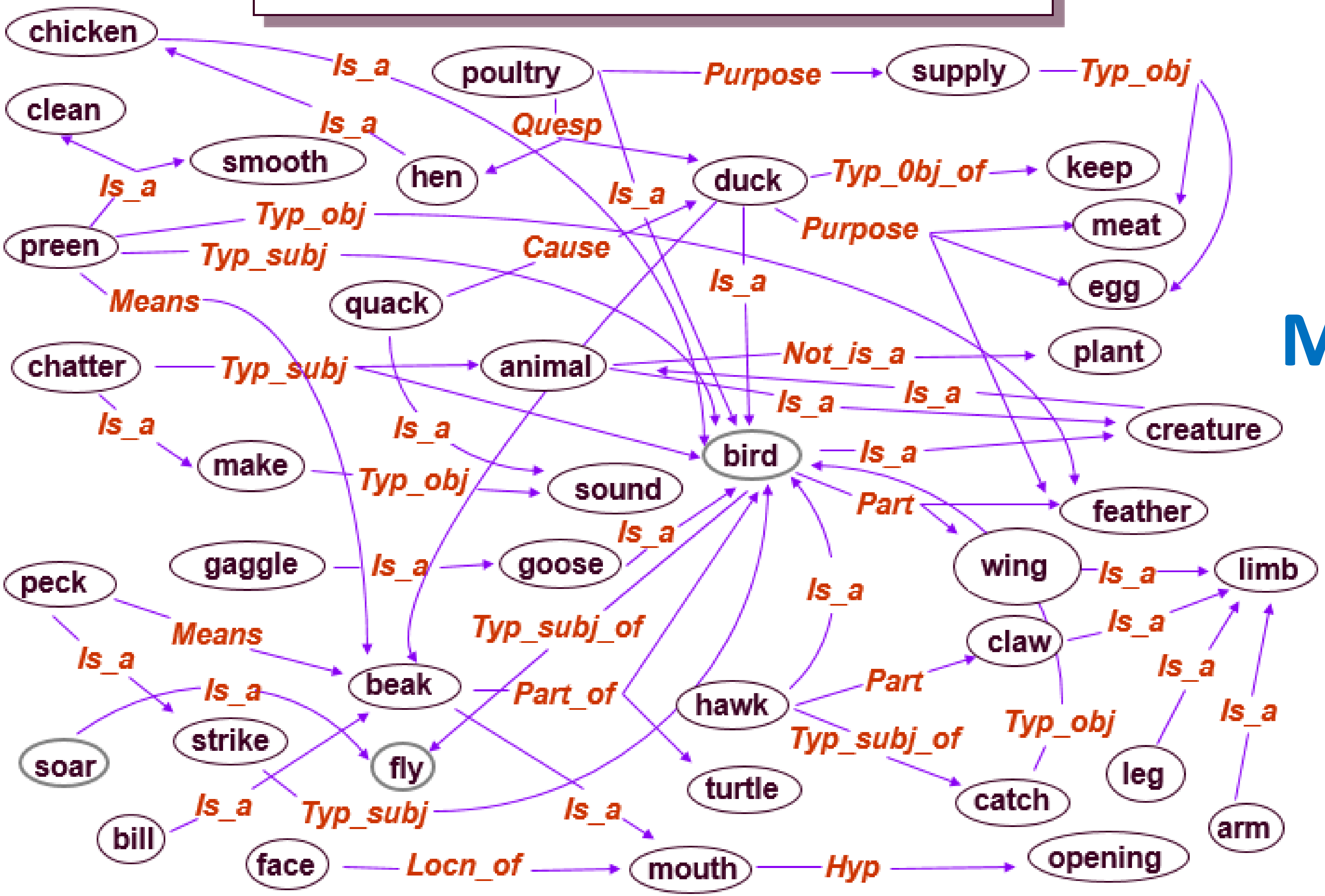
- Symbolic approaches to QA
  - Knowledge representation and search in a symbolic space
  - A case study of MSR MindNet
- Neural approaches to MRC and QA
  - Knowledge representation and search in a neural space
  - A case study of ReasoNet
- MS MARCO

Symbolic approaches to QA: product systems

- Production rules
  - condition—action pairs
  - Represent (world) knowledge as a graph
- Working memory
  - Contains a description of the current state of the world in a reasoning process
- Recongizer-act controller
  - Update working memory by searching and firing a production rule

- A case study: MSR MindNet [Dolan+ 93; Richardson+ 98]

Fragment of lexical space surrounding "bird"



MindNet

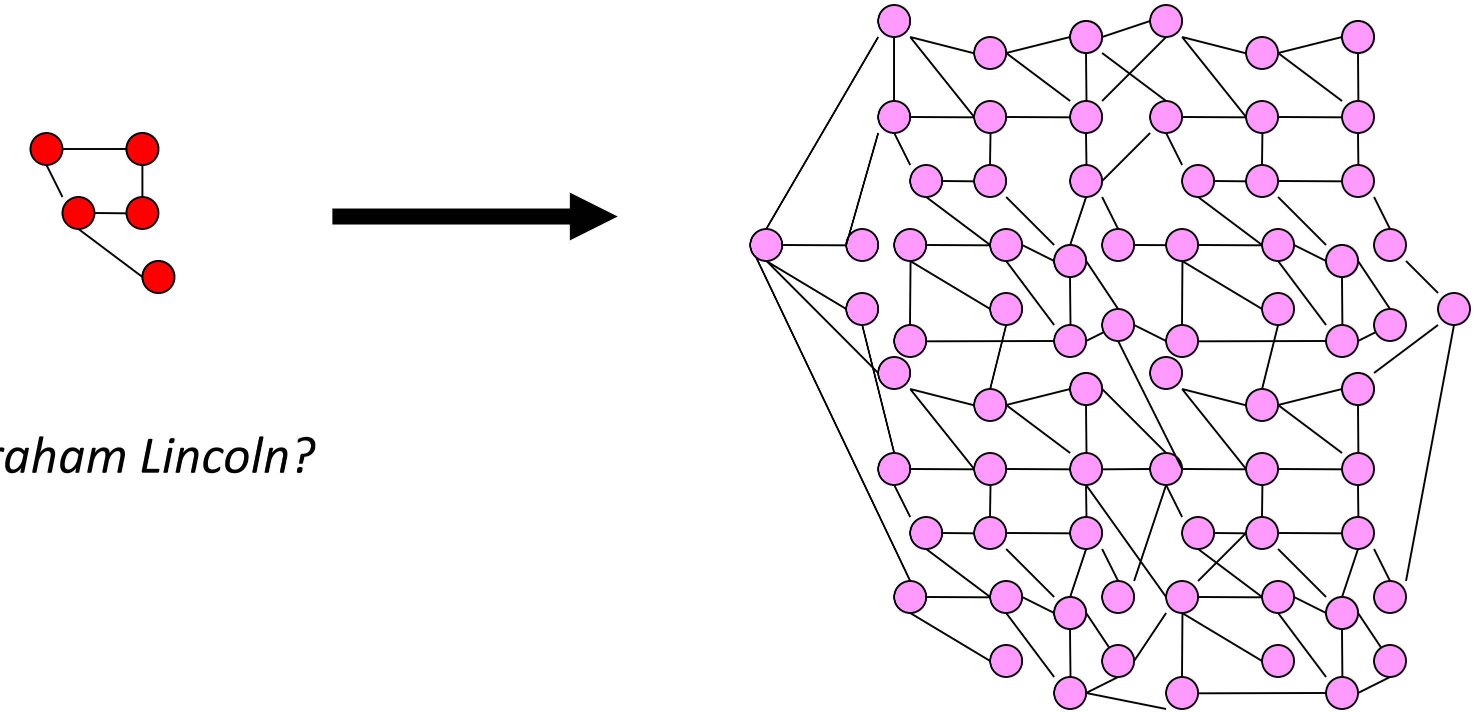
Question Answering with MindNet

- Build a MindNet graph from:
  - Text of dictionaries
  - Target corpus, e.g. an encyclopedia (Encarta 98)
- Build a dependency graph from query
- Model QA as a graph matching procedure
  - Heuristic fuzzy matching for synonyms, named entities, wh-words, etc.
  - Some common sense reasoning (e.g. dates, math)
- Generate answer string from matched subgraph
  - Including well-formed answers that didn't occur in original corpus

Logical Form Matching

Input LF:

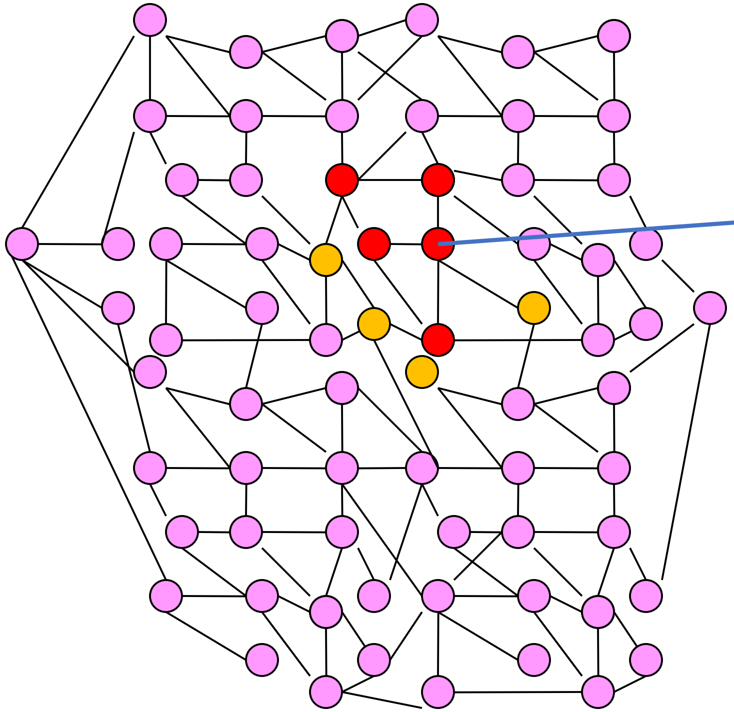
MindNet

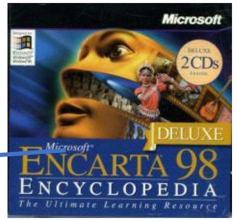


Who assassinated Abraham Lincoln?

“You shall know a word by the company it keeps” (Firth, 1957)

Fuzzy Match against MindNet

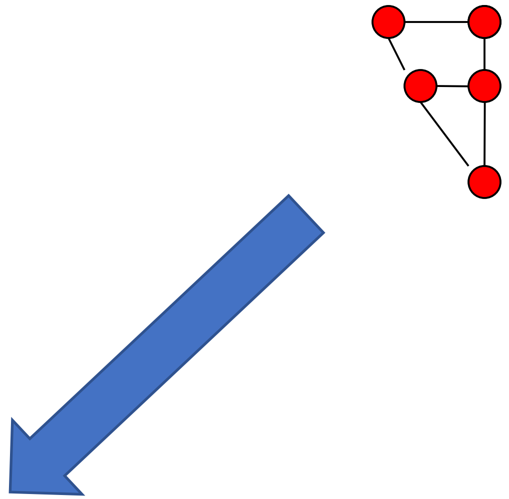




<u>Lincoln, Abraham</u>

American actor <u>John Wilkes Booth</u>, who was a violent backer of the South during the Civil War, <u>shot</u> <u>Abraham Lincoln</u> at Ford's Theater in Washington, D.C., on April 14, 1865.

Generate output string



"John Wilkes Booth shot Abraham Lincoln"

Worked beautifully!

- Just not very often…
- Most of the time, the approach failed to produce any answer at all, even when:
  - An exact answer was present in the target corpus
  - Linguistic analysis for query/target strings was correct
- What went wrong?
  - One major reason: paraphrase alternations

Keyword passage retrieval outperformed all that clever NLP/AI machinery

Example: “How long is the X river?”

- The Mississippi River is 3,734 km (2,320 mi) long.
- …is nearly 86 km long…
- ...is a short river, some 4.5 miles (7.2 km) in length
- The total length of the river is 2,145 kilometres (1,333 mi).
- … at the estimated length of 5,464 km (3,395 mi)…
- **…**is a 25-mile (40 km) tributary of …
- … has a meander length of 444 miles (715 km)…
- … Bali’s longest river, measuring approximately 75 kilometers from source to mouth.
- The … mainstem is 2.75 miles (4.43 km) long although total distance from headwater source tributaries to the sea is 14 miles (23 km).

- …is 314 km long
- …is nearly 86 km long…
- … is a 92-mile (148 km) long tributary of the…
- ...is a short river, some 4.5 miles (7.2 km) in length
- …flows nearly 20 miles (32 km) to the west
- The [river], which is 6,853 km (4,258 miles) long…
- It runs a course of about 105 kilometers
- The 1,450-mile-long (2,330 km) [river] drains…
- ...a 234-mile (377-kilometer) man-made waterway…
- … at the estimated length of 5,464 km (3,395 mi)…
- ... stretches for 2,639 miles (4,247 km).
- …is a 25-mile (40 km) tributary of …
- …starting in and flowing for nearly 160 kilometers through….
- …flows almost 70 stream miles.
- The river runs 184 kilometers before joining…
- … Bali's longest river, measuring approximately 75 kilometers from source to mouth.
- …is reported to be anywhere from 5,499 to 6,690 kilometres (3,417 to 4,157 mi). Often it is said to be "about" 6,650 kilometres (4,130 mi) long.
- ...reaches a length of approximately 25 kilometres
- The length of the Ouse alone is about 52 miles (84 km).
- Measuring a length of 60 kilometers, the [river] flows through
- It has a total length of 925 km (575 mi).
- The total length of the river is 2,145 kilometres (1,333 mi).
- Its length is 209 km…
- …is about 1,180 miles (1,900 km) in length.
- ...the river flows for more than 1,200 km (750 mi)
- …the river proper flows only for 113 km…
- …flows slowly for 900 kilometres (560 mi)…
- … has a meander length of 444 miles (715 km)…
- …is a 350-kilometre (220 mi) long river in …
- it …meanders slowly southwards for 2,320 miles (3,730 km) to …
- The river's main stem is about 71 miles (114 km) long. Its length to its most distant headwater tributary is about 220 miles (350 km).
- After approximately 30 kilometres (19 mi) of its 78-kilometre (48 mi) course, it ….
- …is the longest river in the United Kingdom, at about 220 miles (354 km).
- … is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)…
- The … mainstem is 2.75 miles (4.43 km) long although total distance from headwater source tributaries to the sea is 14 miles (23 km).
- At 320 kilometres (200 mi) (with some estimates ranging up to 596 kilometres (370 mi))...

Back to today, 20 years later…

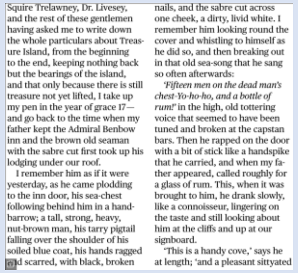
- We're still far from "understanding"
- But we've made great progress!
  - Bigger data, better hardware
  - **Better Algos, esp. neural networks, Deep Learning, Reinforcement Learning…**

- Same fundamental viewpoint

  "You shall know a word by the company it keeps"

Symbolic Space

- **Knowledge Representation**
  - *Explicitly* store a BIG but incomplete knowledge graph (KG)
  - Words, relations, templates
  - High-dim, discrete, sparse vectors
- **Inference**
  - Slow on a big KG
  - Keyword/template matching is sensitive to paraphrase alternations
- **Human comprehensible but not computationally efficient**

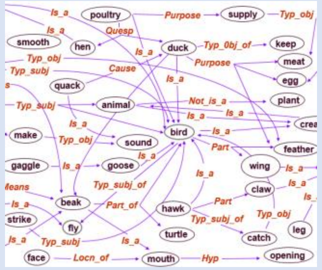


Squire Trelawney, Dr. Livesey, and the rest of these gentlemen having asked me to write down the whole particulars about Treasure Island, from the beginning to the end, keeping nothing back but the bearings of the island, and that only because there is still treasure not yet lifted, I take up my pen in the year of grace 17— and go back to the time when my father kept the Admiral Benbow inn and the brown old seaman with the sabre cut first took up his lodging under our roof.

I remember him as if it were yesterday, as he came plodding to the inn door, his sea-chest following behind him in a hand-barrow; a tall, strong, heavy, nut-brown man, his tarry pigtail falling over the shoulder of his soiled blue coat, his hands ragged and scarred, with black, broken nails, and the sabre cut across one cheek, a dirty, livid white. I remember him looking round the cover and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

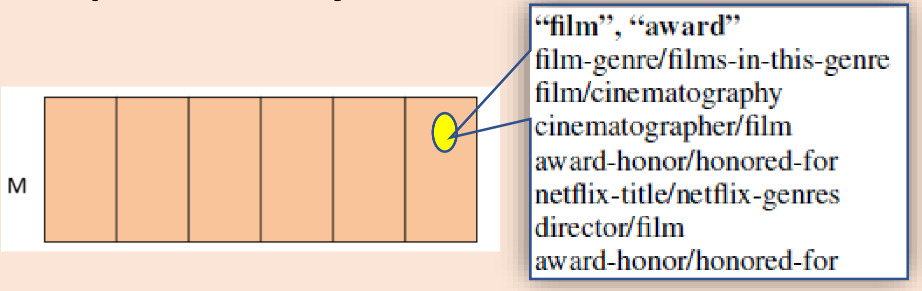
'*Fifteen men on the dead man's chest-Yo-ho-ho, and a bottle of rum!*' in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was brought to him, he drank slowly, like a connoisseur, lingering on the taste and still looking about him at the cliffs and up at our signboard.

'This is a handy cove,' says he at length; 'and a pleasant sittyated



Neural Space

- **Knowledge Representation**
  - *Implicitly* store entities and structure of KG in a *compact* way that is more generalizable
  - Semantic concepts/classes
  - Low-dim, cont., dense vectors shaped by KG
- **Inference**
  - Fast on compact memory
  - Semantic matching is robust to paraphrase alternations
- **Computationally efficient but not human comprehensible *yet***



From symbolic to neural computation

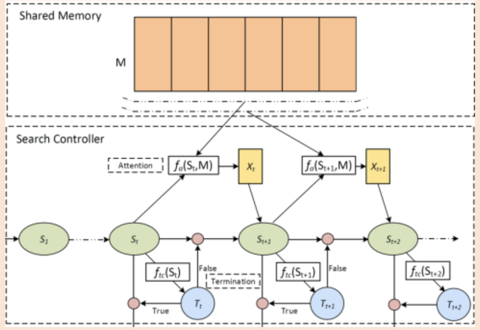
Input: Q

Question: Symbolic → Neural
by embedding models / encoder

Symbolic Space
- UI: human readable I/O
- Leverage traditional symbolic approaches as pre/post processing
    - Keyword matching
    - Ontology based models
    - e.g., doc/passage/entity search/ranking

Inference: Question + KG → Answer
by RNN + Attention

Neural Space

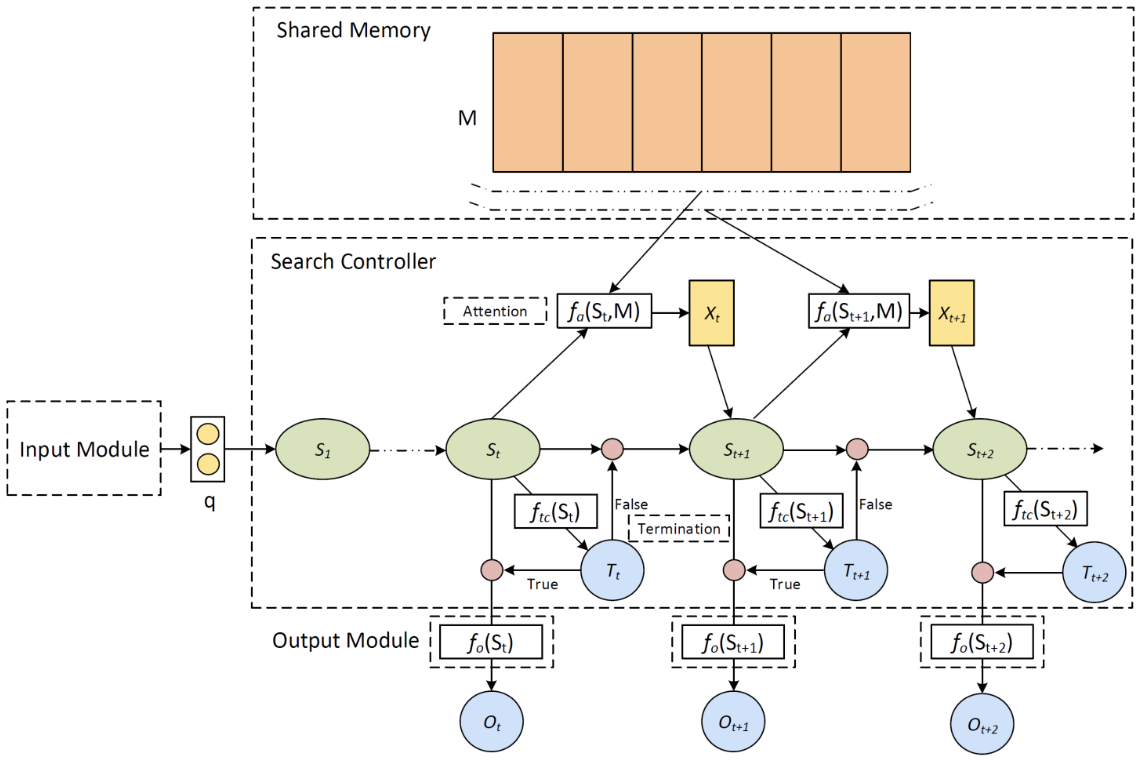


Error(A, A*)

Output: A

Answer: Neural → Symbolic
by generative models / decoder

Case study: ReasoNet with Shared Memory



- **Production Rules** → **Shared memory** encodes task-specific knowledge
- **Working memory** → **Hidden state S_t** Contains a description of the current state of the world in a reasoning process
- **Recognizer-act controller** → **Search controller** performs multi-step inference to update S_t of a question using knowledge in shared memory

- Shared memory and search controller are jointly learned via SL+RL
- Input/output modules are task-specific

[Shen+ 16a]

QA on Text

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and **Von Miller** for Denver.
Answer	**Von Miller**

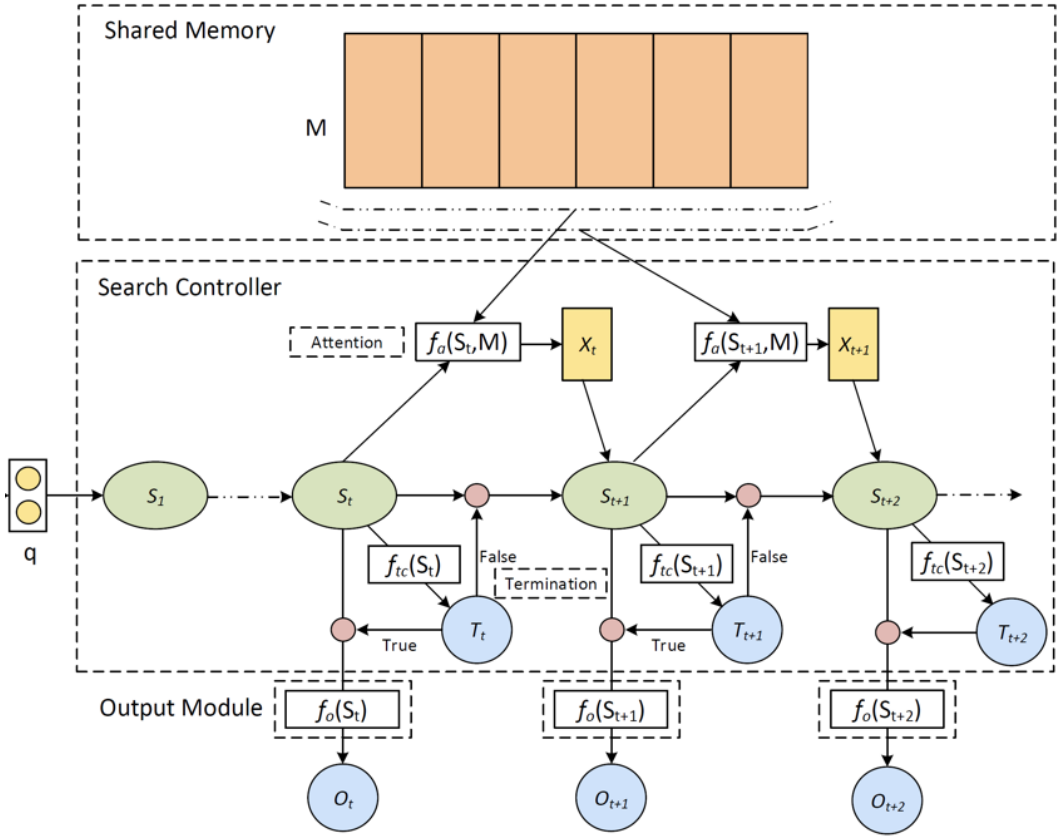
Multi-step inference:

- Step 1:
  - **Extract:** Manning is #1 pick of 1998
  - **Infer:** Manning is NOT the answer
- Step 2:
  - **Extract:** Newton is #1 pick of 2011
  - **Infer:** Newton is NOT the answer
- Step 3:
  - **Extract:** Newton and Von Miller are top 2 picks of 2011
  - **Infer:** **Von Miller** is the #2 pick of 2011

ReasoNet: Learn to Stop Reading

Keep gathering information (encoded in internal state) until a good answer is formed

1. Given a set of docs in memory: $\mathbf{M}$
2. Start with query: S
3. Identify info in $\mathbf{M}$ that is related to S : $X = f_a(S, \mathbf{M})$
4. Update internal state: $S = \mathrm{RNN}(S, X)$
5. Whether a satisfied answer O can be formed based on S: $f_{tc}(S)$
6. If so, stop and output answer $O = f_o(S)$; otherwise return to 3.



[Shen+ 16b]

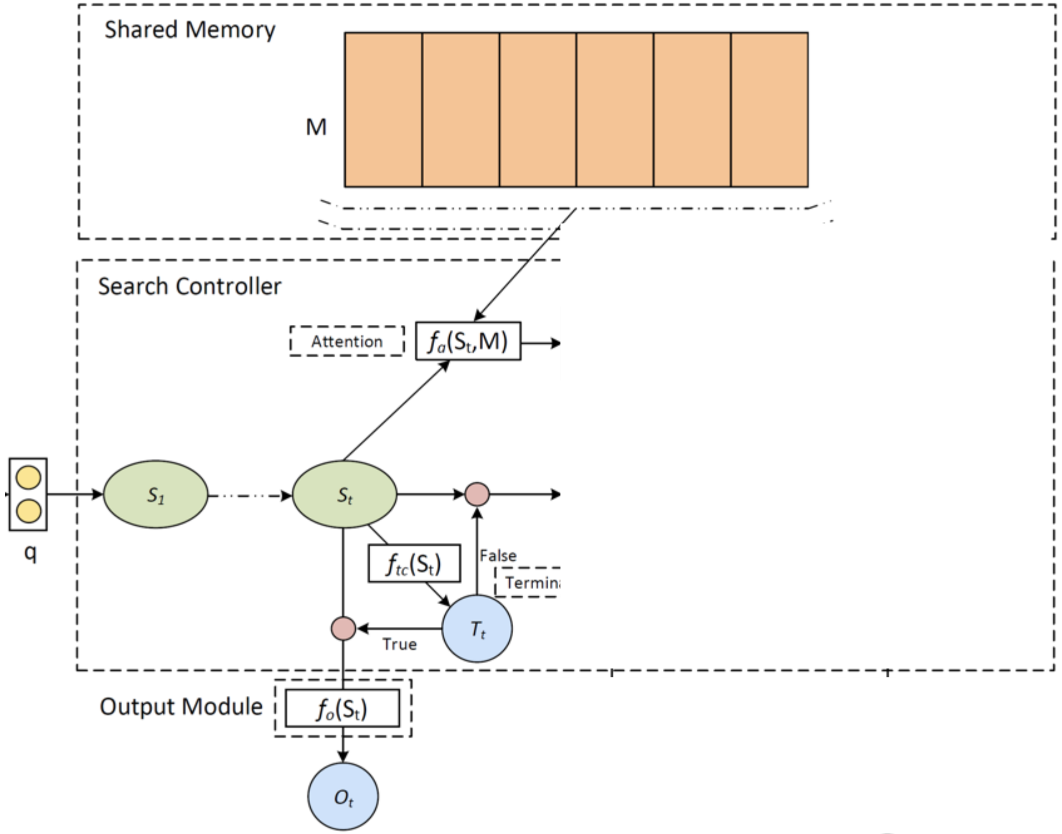
ReasoNet at work

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and Von Miller for Denver.
Answer	**Von Miller**

Rank-1

Rank-2

Rank-3

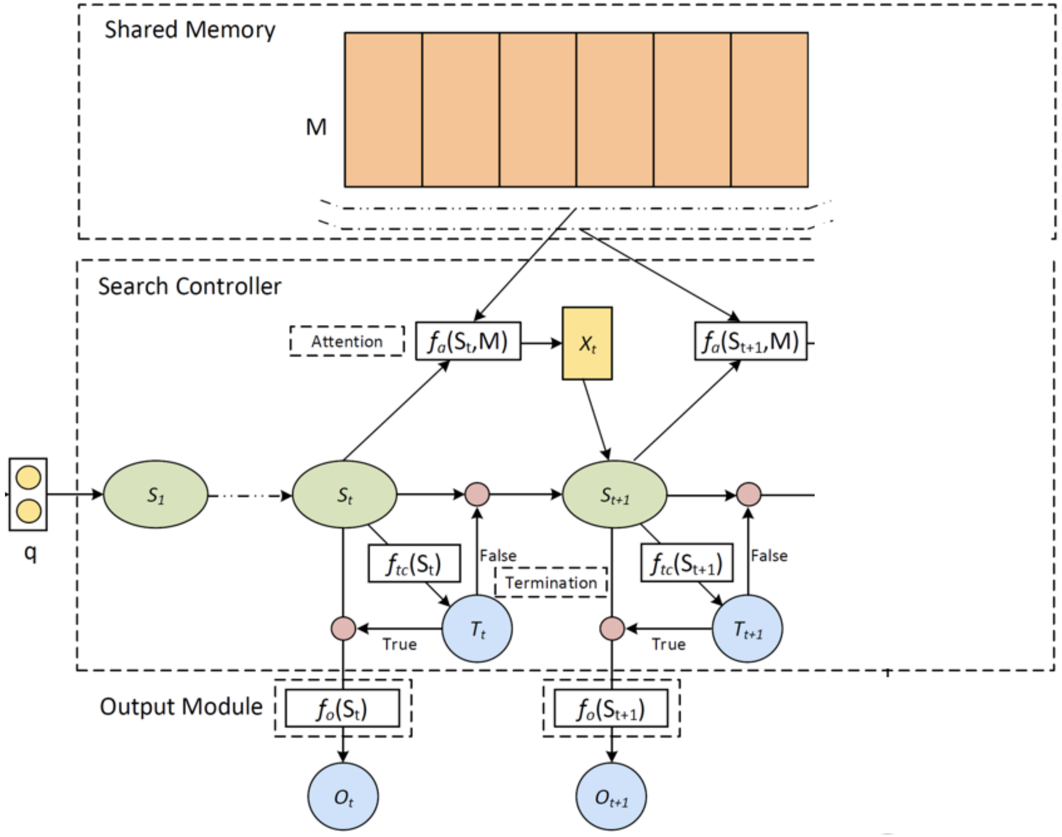


Step	Termination Probability	Prob. Answer
1	0.001	0.392

ReasoNet at work

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and Von Miller for Denver.
Answer	**Von Miller**

Rank-1
Rank-2
Rank-3

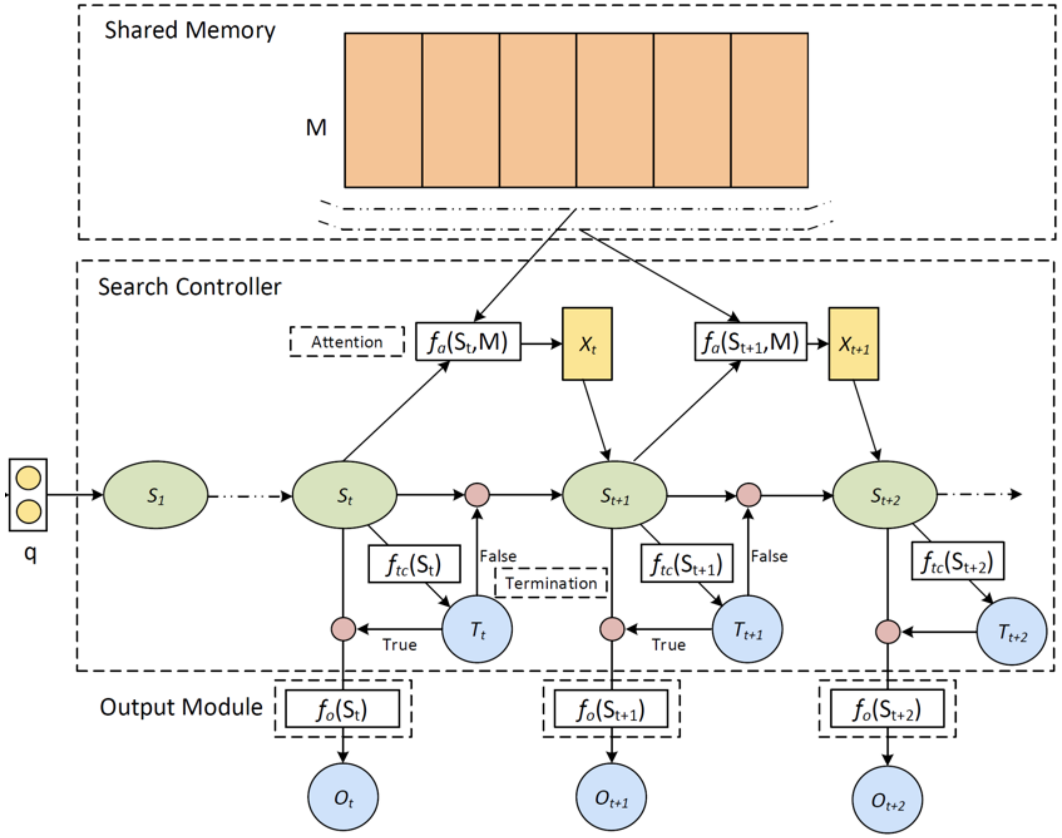


Step	Termination Probability	Prob. Answer
1	0.001	0.392
2	0.675	0.649

ReasoNet at work

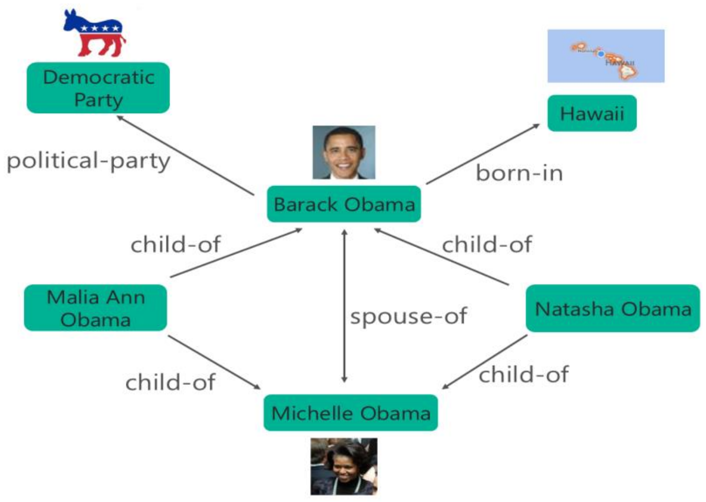
Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and Von Miller for Denver.
Answer	**Von Miller**

Rank-1
Rank-2
Rank-3



Step t	Termination Probability f_{tc}	Prob. Answer f_o
1	0.001	0.392
2	0.675	0.649
3	0.939	0.865

Question Answering (QA) on Knowledge Base



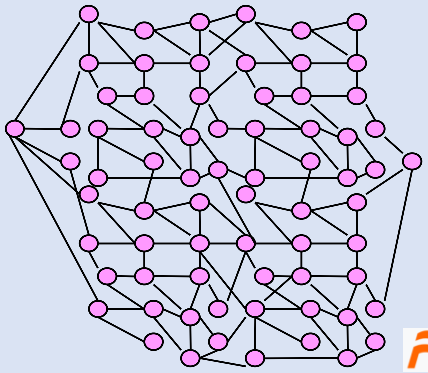
Large-scale knowledge graphs
- Properties of billions of entities
- Plus relations among them

An QA Example:

Question: `what is Obama's citizenship?`
- Query parsing:
  `(Obama, Citizenship,?)`
- Identify and infer over relevant subgraphs:
  `(Obama, BornIn, Hawaii)`
  `(Hawaii, PartOf, USA)`
- correlating semantically relevant relations:
  `BornIn ~ Citizenship`

Answer: `USA`

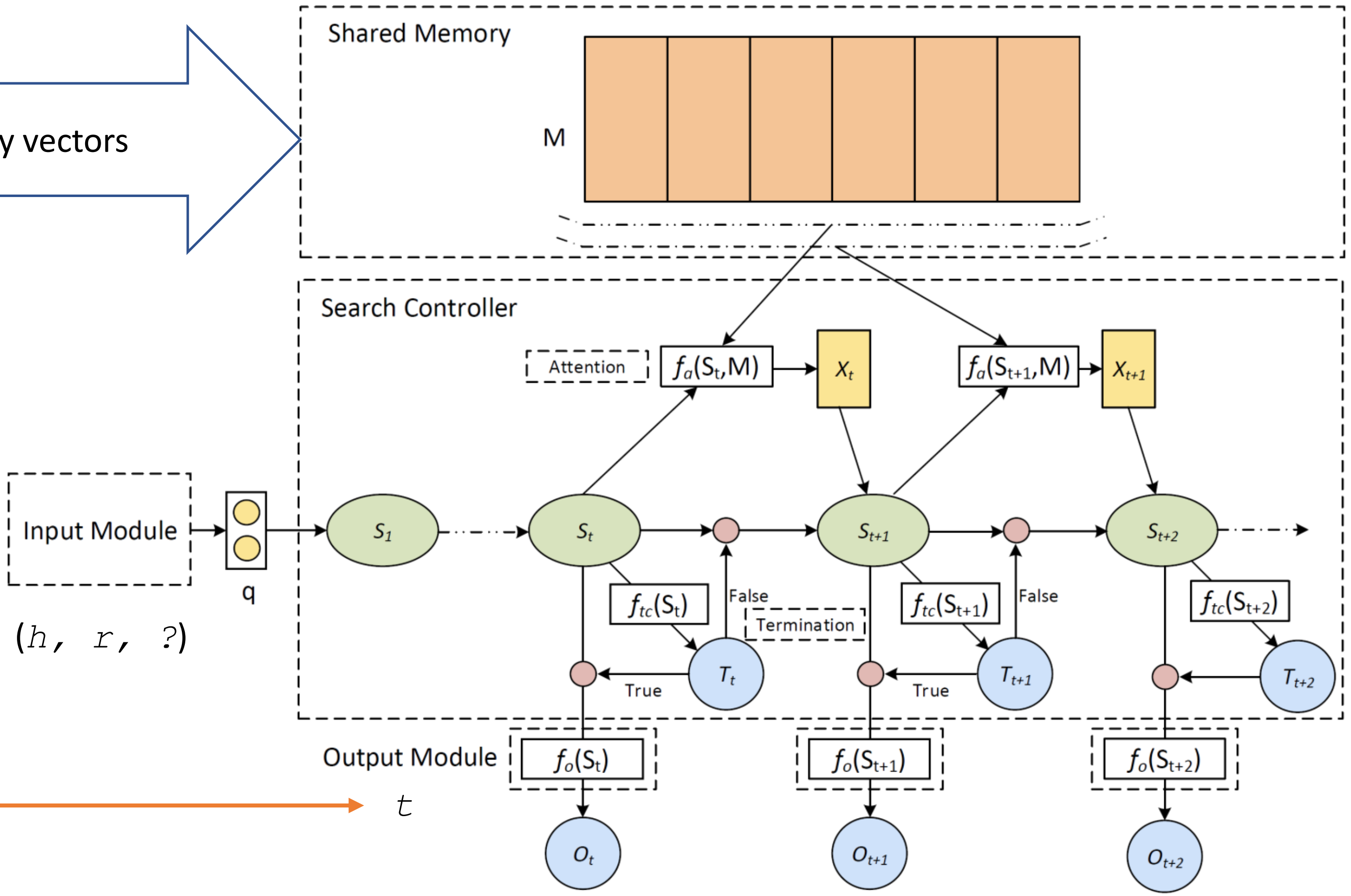
ReasoNet: joint learning of Shared Memory and Search Controller



Embed KG to memory vectors

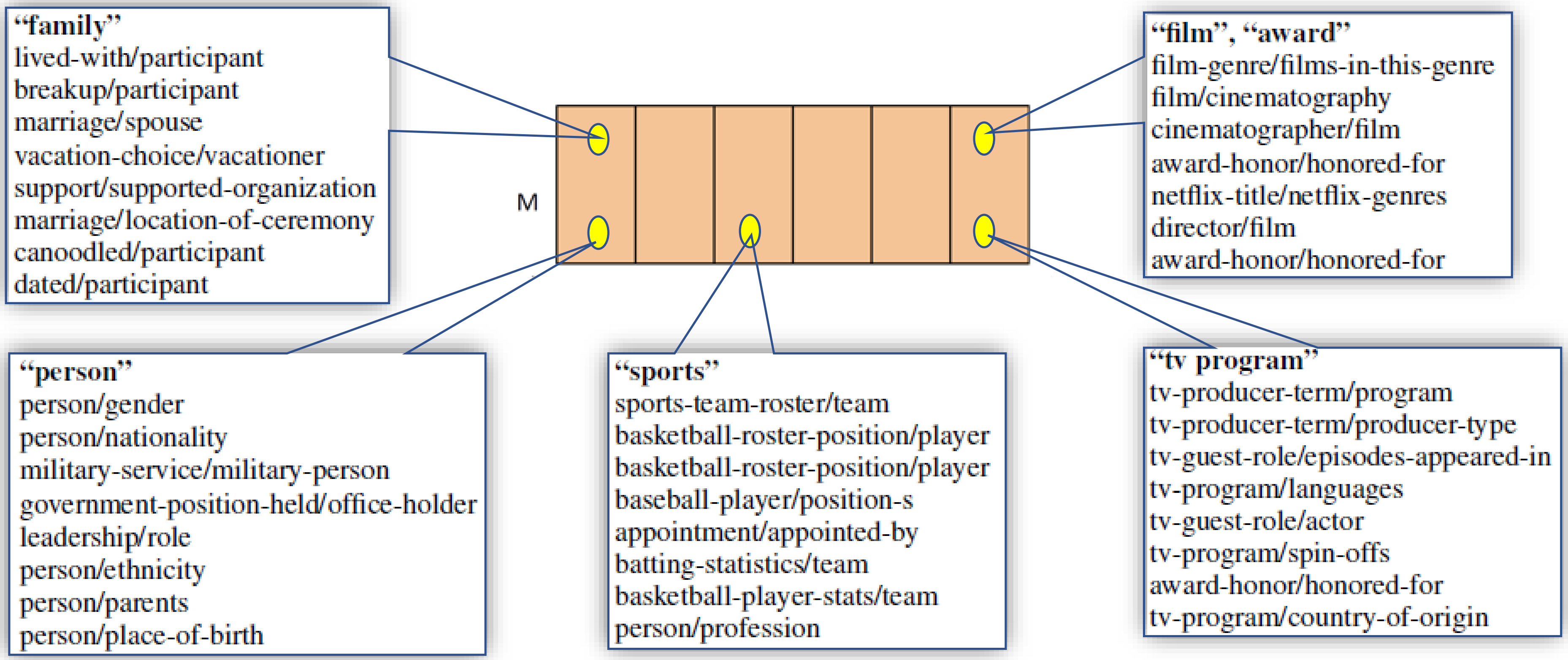
Training samples from KG:

```
(Obama, BornIn, Hawaii)
(Hawaii, PartOf, USA)
…
(h, r, t)
…
(Obama, Citizenship,?)->(USA)
```



Shared Memory: long-term memory to store learned knowledge, like human brain

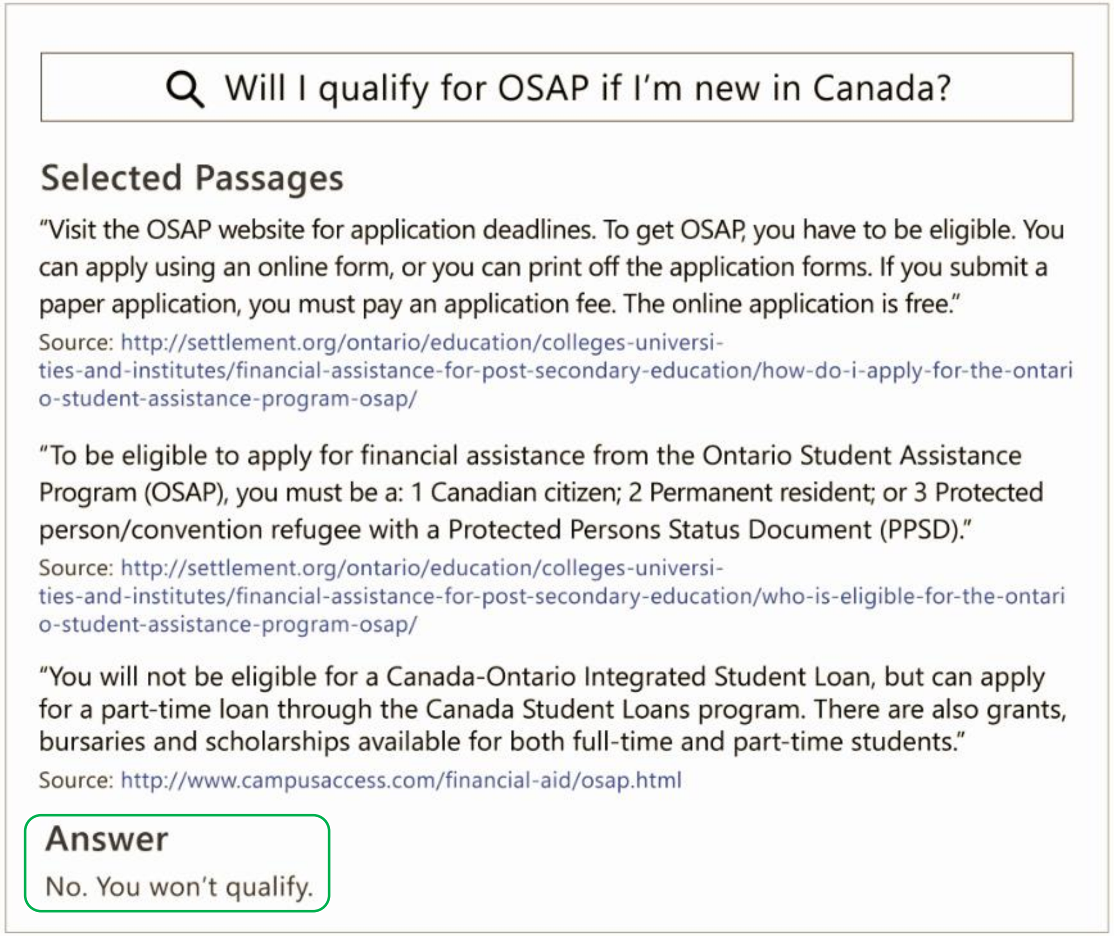
- Knowledge is learned via performing tasks, e.g., update memory to answer new questions
- New knowledge is *implicitly* stored in memory cells via gradient update
- Semantically relevant relations/entities can be compactly represented using similar vectors.



MS MARCO

Evolution of MRC Datasets

Dataset	Company	Query Source	Answer	# Queries	# Docs
MC Test	Microsoft	Crowdsourced	Multiple Choice	2640	660
WikiQA	Microsoft	User Logs	Sentence Selection	3047	29K sentences
CNN/DailyMail	DeepMind	Cloze	Fill in entity	1.4M	93K CNN, 220K DM
Children's Book	Facebook	Cloze	Fill in the word	688K	688K contexts
SQuAD	Stanford	Crowdsourced	Span	100K	536
News QA	Maluuba	Crowdsourced	Span	120K	12K
MS MARCO	Microsoft	User Logs	Human Synthesized	100k	1M passages, 200K+ docs



A Thorough Examination of the CNN/DailyMail Reading Comprehension Task

MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/MCTest_EMNLP2013.pdf

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

About MS MARCO

- **M**icro**s**oft **M**achine **R**eading **Co**mprehension **D**ataset
- What’s unique about it?
  - Real world questions
  - Real world documents
  - Human generated answer
  - Some questions are unanswerable
  - Some answers need inference: Yes/No answers

Examples

Refining the Evidence Snippet
Q: what are the different types of skin ulcers
P: This medical symptom information shows the various types of Skin ulcer, and other related symptoms or conditions, including their causes and diagnosis. 1 **Pressure sore**. 2 **Decubitus ulcers**. 3 **Weals**.
A: Pressure sore, Decubitus ulcers, and Weals.
Combining Evidence with Question Words
Q: what is the economic impact of the japan earthquake
P: Impact on Japan's Economy. The Triple Disaster devastated Japan's economy in four ways. It destroyed 138,000 buildings and cost **$360 billion in economic damage**.
A: Japan have $ 360 billion economic damage because of earthquake.
Multiple Evidence from One Passage
Q: what is the time span of the roman empire
P: The history of the Roman Empire covers the history of Ancient Rome from the fall of the Roman Republic in **27 BC** until the abdication of the last Emperor in **476 AD**.
A: 27 BC to 476 AD
Multiple Evidence from Multiple Passages
Q: who did odysseus see in the underworld
P_1: The souls that Odysseus saw in the Underworld On seeing **Achilles**' soul, said Odysseus: Achilles, the most fortunate man that ever was or will be honored as though you were a god and now you are a mighty prince among the dead.
P_2: Odysseus talked to his mother Anticlea, who died of grief when he did not return home after the Trojan War. Odysseus was also surprised to see **Elphenor**, the youngest member of his crew, in the Underworld.
A: Elphenor and Achilles.
Inference
Q: is there an age limit for learning speech
P: Age is not a detriment to language learning, and by all accounts, learning a second (or third etc) language actually keeps the older language learners mind active. People of all ages can benefit from learning languages.
A: No

Category
Exact Match (68.06%)
Refining the Evidence Snippet (21.30%)
Combining Question Words (2.30%)
Multiple Evidence (3.85%)
Inference (4.48%)

S-NET: FROM ANSWER EXTRACTION TO ANSWER GENERATION FOR MACHINE READING COMPREHENSION

https://arxiv.org/pdf/1706.04815.pdf

Summary

- Symbolic approaches to QA
  - Knowledge representation and search in a symbolic space
  - A case study of MSR MindNet
- Neural approaches to MRC and QA
  - Knowledge representation and search in a neural space
  - A case study of ReasoNet
- Ongoing research
  - New MRC dataset, MS MARCO
  - Neural approaches to symbolic reasoning
  - Interpret or visualize the reasoning process in neural space

Thank you

Microsoft