Survey of Deep, Vector-based Models for Question Answering

Corbin Rosset

Under supervision of Drs. Raman Arora and Mark Dredze

No content is mine

18 September 2016

QA task and available data

In research settings, the QA task is conventionally reduced to a classification task over small number of entities, or predict missing entity in a KB triple.

Inputs: a series of sentences, documents, and/or perhaps a knowledge base, a question, and at train time – an answer
Outputs: a single word, noun phrase, or sentence as an answer to an unambiguous question.
Datasets: (KB): ReVerb, FreeBase, DBpedia; (Paraphrase): WikiAnswers; (QA): WebQuestions

Types of QA tasks:

1. Yes/No
2. One-shot Factoid – a single query whose answer type is an entity, known to be public information
    1. **Single-relation** (Who was Kennedy's Vice President?)
    2. **Double-relation** (Who was Vice President after Kennedy died?)
3. Descriptive – single query answered by a description of a phenomenon or entity
4. Multi-turn Factoid – multiple queries around a central topic, all answered by publicly known information.
5. Multi-turn Situational – multiple queries around a *personal* situation – public knowledge is not useful

Examples

Google who is the president of France

All News Images Shopping Videos More Search tools

About 314,000,000 results (0.84 seconds)

France / President

François Hollande

More about François Hollande

Feedback

People also ask

- Who was the first president of France?
- How long has Francois Hollande been president?
- Is Hollande a Socialist?

Google what happened at chernobyl

All Videos Images News Maps More Search tools

About 505,000 results (0.58 seconds)

On April 26, 1986, the world's worst nuclear accident **happened** at the **Chernobyl** plant near Pripyat, Ukraine, in the Soviet Union. An explosion and fire in the No. 4 reactor sent radioactivity into the atmosphere. Apr 26, 2012

April 26, 1986 | Explosion at Chernobyl Nuclear Plant - The New York ...
learning.blogs.nytimes.com/.../april-26-1986-explos... New York Times Learning Network

About this result • Feedback

Google how to do laundry

All Videos Images Shopping Books More Search tools

About 40,400,000 results (0.49 seconds)

1. Intro: How to Do Your Own Laundry (basic) Time to complete: Approximately 45 mins. ...
2. Step 1: Separate Clothes. Separate the clothes into piles. ...
3. Step 2: Use Dial to Start Washing Machine. Turn the dial to set the washing type. ...
4. Step 3: Load Machine. ...
5. Step 4: Add Detergent. ...
6. Step 5: Close lid. ...
7. Step 6: Wait.

How to Do Your Own Laundry (basic) - Instructables
www.instructables.com/id/**How-to-Do**-Your-Own-**Laundry**-basic/ Instructables

About this result • Feedback

Facebook's BaBi Dataset

"Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks", Jason Weston, Antione Bordes, Tomas Mikolov

20 proxy tasks to evaluate comprehension via question answering. Goal is clearly identify limitations of a system

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions
John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

Task 11: Basic Coreference
Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel? A:studio

Task 12: Conjunction
Mary and Jeff went to the kitchen.
Then Jeff went to the park.
Where is Mary? A: kitchen
Where is Jeff? A: park

Task 13: Compound Coreference
Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning
In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Task 15: Basic Deduction
Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

Task 16: Basic Induction
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg? A:white

Task 17: Positional Reasoning
The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red sphere to the right of the blue square? A:yes
Is the red square to the left of the triangle? A:yes

Task 18: Size Reasoning
The football fits in the suitcase.
The suitcase fits in the cupboard.
The box is smaller than the football.
Will the box fit in the suitcase? A:yes
Will the cupboard fit in the box? A:no

Task 19: Path Finding
The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

Task 20: Agent's Motivations
John is hungry.
John goes to the kitchen.
John grabbed the apple there.
Daniel is hungry.
Where does Daniel go? A:kitchen
Why did John go to the kitchen? A:hungry

Facebook's BaBi Dataset

"Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks", Jason Weston, Antione Bordes, Tomas Mikolov

Some tasks require significant comprehension of semantics, but reasoning is simple; other require complex reasoning over multiple steps, but little semantic understanding.

Three main sources of error: imperfect semantic understanding, inadequate knowledge encoding, or insufficient model capacity.

Pathfinding and positional reasoning are notoriously difficult

Task I: path finding

```
1.The office is east of the hallway.
2.The kitchen is north of the office.
3.The garden is west of the bedroom.
4.The office is west of the garden.
5.The bathroom is north of the garden.
How do you go from the kitchen to the garden?  south, east, relies on 2 and 4
How do you go from the office to the bathroom?  east, north, relies on 4 and 5
```

Task II: positional reasoning

```
1.The triangle is above the pink rectangle.
2.The blue square is to the left of the triangle.
Is the pink rectangle to the right of the blue square?  Yes, relies on 1 and 2
Is the blue square below the pink rectangle?  No, relies on 1 and 2
```

Two Kinds of Algorithms for Open Domain QA

Two flavors of algorithms: **information retrieval** based, and **knowledge base**:

- KB: a store of triples (subject, relationship, object) represented as a graph i.e. (*cantonese, spoken-in, Hong Kong*)
  - *e.g.*: Single-relation queries are a triple with a missing entity, answer is the entity
  - *Task*: Map natural language questions to semantic representation for an existing DB/KB
  - *Subtask*: entity linking (string from NL query to entity in KB)
- Information Retreival based: given access to the internet's documents…
  - Given a natural language query, rank and return paragraphs/sentences/phrases in the documents

Information Retrieval-based **pipeline**:

1. Question Processing: parse and detect named entities; infer the answer type and focus of the sentence
2. Query Reformulation: expand and format query for appropriate KB, DB, or IR engine
3. Document Retrieval: retrieve ranked documents, then identify, filter, and re-rank relevant paragraphs
4. Answer Processing: extract and rank candidate answers from relevant paragraphs

Other challenges:

- Encode ordering of words in sentences (*What are cats afraid of*? Vs *What's afraid of cats?*)
- Encode ordering of supporting facts (*Sally put the ball in the box; then she put the box in the kitchen*)
- Encode multiple KB triples into a single vector/as a tensor product?
- Flexible coverage of paraphrasic queries

Traditional Algorithms for QA

JOHNS HOPKINS
UNIVERSITY

18 September 2016

Level of Human Intervention

Speech and Language Processing, Jurafsky and Martin. Chapter 28 pg 5.

Right: Question Topology from Li and Roth 2002 "Learning Question Classifiers". A corpus of 5500 labeled questions with course or fine-grained tags.

Below: semantic parser output of natural language text for various structured database formats.

In traditional systems, much manual labor goes into feature engineering, Grammar/Regex constructions, and hierarchical tag/labels. These are difficult to maintain and do not generalize well, especially as queries become more complex.

Question	Logical form
When was Ada Lovelace born?	`birth-year (Ada Lovelace, ?x)`
What states border Texas?	λ x.state(x) $\wedge$ borders(x,texas)
What is the largest state	argmax(λx.state(x), λx.size(x))
How many people survived the sinking of the Titanic	`(count (!fb:event.disaster.survivors fb:en.sinking_of_the_titanic))`

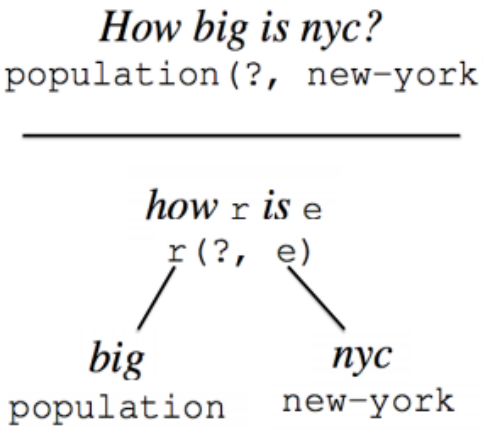
Tag	Example
ABBREVIATION	
abb	What's the abbreviation for limited partnership?
exp	What does the "c" stand for in the equation E=mc2?
DESCRIPTION	
definition	What are tannins?
description	What are the words to the Canadian National anthem?
manner	How can you get rust stains out of clothing?
reason	What caused the Titanic to sink ?
ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?
HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

Weakly supervised QA (1): Paralex

"Paraphrase-Driven Learning for Open Question Answering" Fader, Zettlemoyer, Etzioni (2013)

One attempt to address lack of coverage/generalizability of QA systems to understanding different wordings of the same query: paraphrases

- Learn a lexical map from NL questions to single-relation triples that ranks the best DB query "templates".
- The templates are expanded from a small seed lexicon using a paraphrase dataset and a word alignment algorithm.
- Metrics used are accuracy, precision, recall, F1 to evaluate correct answers



1) Match the NL question to a form in the lexicon
2) Substitute NL words with DB entities and relationships

String	Learned Database Relations for String
get rid of	treatment-for, cause, get-rid-of, cure-for, easiest-way-to-get-rid-of
word	word-for, slang-term-for, definition-of, meaning-of, synonym-of
speak	speak-language-in, language-speak-in, principal-language-of, dialect-of
useful	main-use-of, purpose-of, importance-of, property-of, usefulness-of

String	Learned Database Entities for String
smoking	smoking, tobacco-smoking, cigarette, smoking-cigar, smoke, quit-smoking
radiation	radiation, electromagnetic-radiation, nuclear-radiation
vancouver	vancouver, vancouver-city, vancouver-island, vancouver-british-columbia
protein	protein, protein-synthesis, plasma-protein, monomer, dna

Weakly supervised QA (2): Embeddings

"Open Question Answering with Weakly Supervised Embedding Models" Bordes, Weston, Usunier (2014)

- One of the first papers to map questions and answers into same embedding space, where the correct answer is most similar to the question embedding using a scoring function $S(q, a) = f(q) * g(a)$
  - Embedding functions, f and g, are linear operators on sparse binary bag-of-words vectors.
- SGD training on margin loss objective (negative examples are generated)
- Paraphrases used to enrich embedding functions: S(q1, q2) where q1 and q2 are labeled paraphrase pairs.
- Fine tuning step: scalar scoring function (dot product) parameterized by similarity matrix
- String matching preprocessing steps are very helpful.

Method	F1	Prec	Recall	MAP
Paralex *(No. 2-arg)*	0.40	**0.86**	0.26	0.12
Paralex	0.54	0. 77	0.42	0.22
Embeddings	0.68	0.68	0.68	0.37
Embeddings *(no paraphrase)*	0.60	0.60	0.60	0.34
Embeddings *(incl. n-grams)*	0.68	0.68	0.68	0.39
Embeddings+fine-tuning	**0.73**	0.73	**0.73**	**0.42**

Performance on re-ranking QA pairs from WikiAnswers + ReVerb test set

KB Triple	Question Pattern
(?, r, e)	*who r e ?*
(?, r, e)	*what r e ?*
(e, r, ?)	*who does e r ?*
(e, r, ?)	*what does e r ?*
(?, r, e)	*what is the r of e ?*
(?, r, e)	*who is the r of e ?*
(e, r, ?)	*what is r by e ?*
(?, r, e)	*who is e's r ?*

KB Triple	Question Pattern
(?, r, e)	*what is e's r ?*
(e, r, ?)	*who is r by e ?*
(e, r-in, ?)	*when did e r ?*
(e, r-on, ?)	*when did e r ?*
(e, r-in, ?)	*when was e r ?*
(e, r-on, ?)	*when was e r ?*
(e, r-in, ?)	*where was e r ?*
(e, r-in, ?)	*where did e r ?*

All questions were generated by selecting a triple from the KB and set of question patterns from the table

Problem: Traditional Representations are only suited for simple queries

single-relational queries into a structured database, simple factoid lookups in tables, etc

JOHNS HOPKINS
UNIVERSITY

Deep Learning Models for QA

Memory Neural Networks and their cousins: architectures with many memory cells have greater capacity for inference

JOHNS HOPKINS
UNIVERSITY

18 September 2016

Memory Neural Networks MemNN

"Memory Networks" Jason Weston, Sumit Chopra, Antoine Bordes (2014)

Use 4 different neural networks to decompose QA into basic steps

1. (Input feature map): any technique to convert text into vector space representation, x
2. (generalization): updates old memories given the new input; compress, rewrite, delete, or generalize its memories. An indexing function H(x) that decides which slot in memory to update.
3. (output feature map): produces a new featuer space output, given the new input and the current memory state. Another scoring function finds the top k most relevant memories for x
4. (response): converts the output into the response format desired. A third scoring function ranks the entities against x and the relevant memories and outputs the best entity. Sentence generation with RNN can also be done here.

1. Convert x to an internal feature representation $I(x)$.
2. Update memories $\mathbf{m}_i$ given the new input: $\mathbf{m}_i = G(\mathbf{m}_i, I(x), \mathbf{m}),\ \forall i.$
3. Compute output features o given the new input and the memory: $o = O(I(x), \mathbf{m})$.
4. Finally, decode output features o to give the final response: $r = R(o)$.

Difficult to train: fully supervised - given desired inputs and responses, and the supporting sentences are labeled

Method	F1
(Fader et al., 2013)	0.54
(Bordes et al., 2014b)	0.73
MemNN (embedding only)	0.72
MemNN (with BoW features)	0.82

Results on Fader et al (2013)'s large sclae QA task

Results of MemNN on new BaBI QA task

"Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks", Jason Weston, Antione Bordes, Tomas Mikolov

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM *COREF+SRL features*	MemNN Weston et al. (2014)	MemNN *ADAPTIVE MEMORY*	MemNN *AM + N-GRAMS*	MemNN *AM + NONLINEAR*	MemNN *AM + NG + NL*	No. of ex. req. $\geq$ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

Positional reasoning and path-finding tasks are most difficult

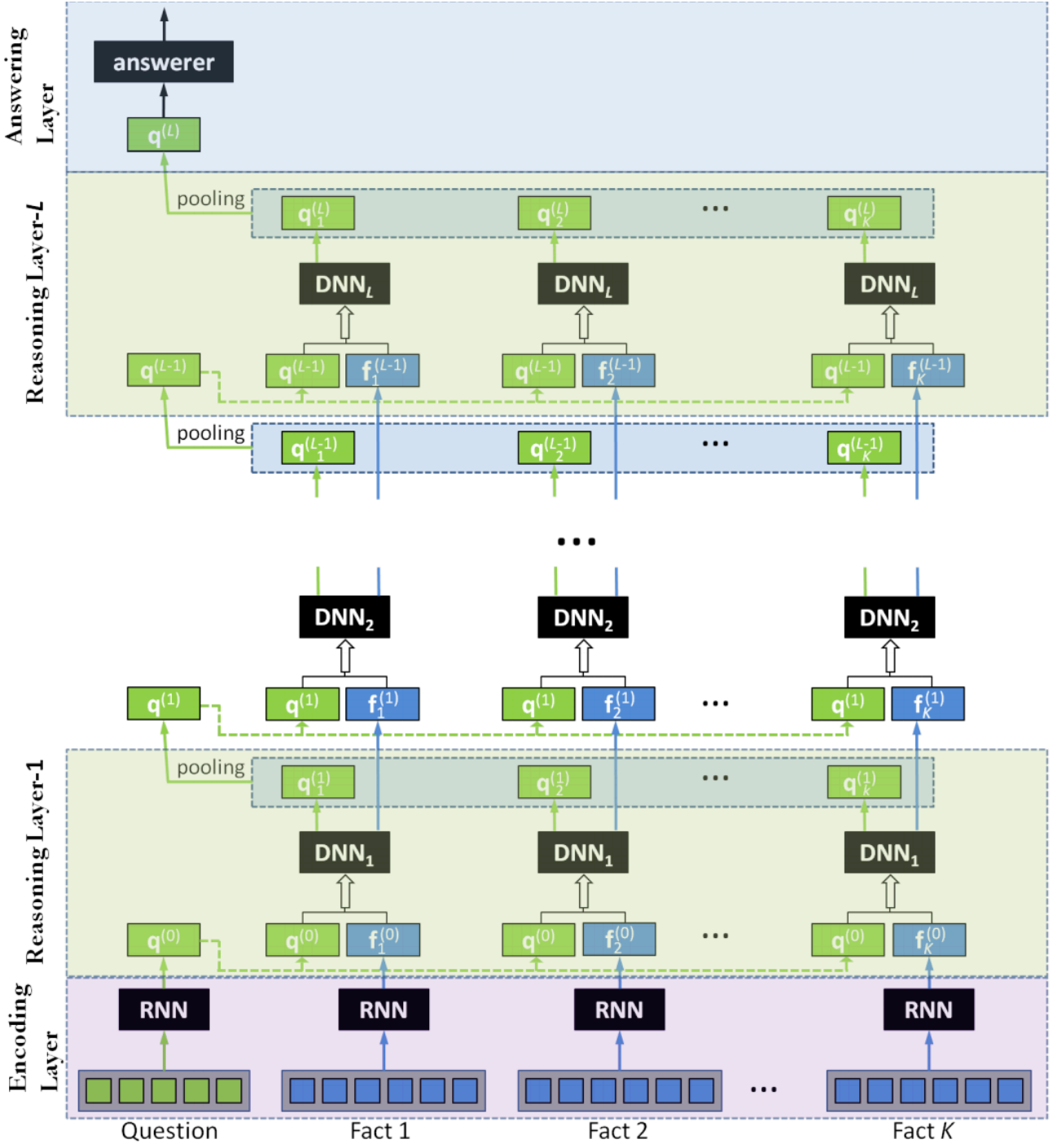
Neural Reasoner

"Towards Neural Network-Based Reasoning" Baolin Peng, Zhengdong Lu, Hang Li, Kam-Fai Wong (2015)

- Iteratively refine representations of the question and supporting facts to capture logical structure
- At each layer, each fact is paired with the question and fed into a neural network to output a new fact representation.
- Deeper in the network, the representations become more abstract.

	Posi. Reason. (1K)	Posi. Reason. (10K)
Step-by-step Supervision		
MEMORY NET-STEP	65.0%	75.4%
DYNAMIC MEMORY NET	59.6%	-
End-to-End		
MEMORY NET-N2N	59.6%	60.3%
NEURAL REASONER	66.4%	97.9%

	Path Finding (1K)	Path Finding (10K)
Step-by-step Supervision		
MEMORY NET-STEP	36.0%	68.1%
DYNAMIC MEMORY NET	34.5%	-
End-to-End		
MEMORY NET-N2N	17.2%	33.4%
NEURAL REASONER	17.3%	87.0%

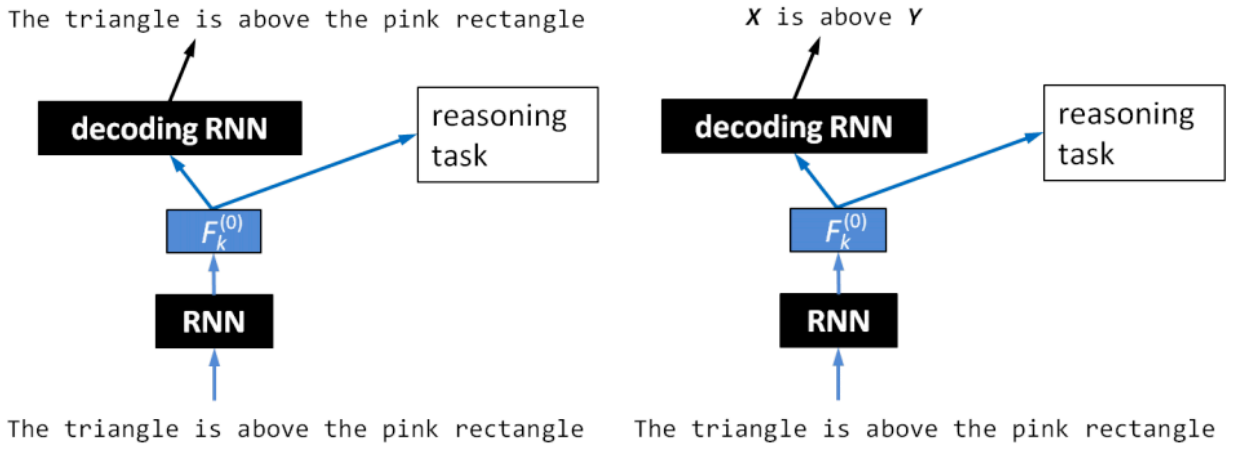


Neural Reasoner – learns abstract representations

"Towards Neural Network-Based Reasoning" Baolin Peng, Zhengdong Lu, Hang Li, Kam-Fai Wong (2015)

- The reasoning task alone cannot give enough supervision for learning accurate word vectors and parameters of the RNN encoder
- **Auxiliary training**: reconstruct the original questions or their more abstract forms with variables to compensate for the lack of supervision in the learning task and introduce beneficial bias
- Define the NN objective as a convex combination of the **classification objective** and the **log-likelihood of the reconstructed question-fact sequence** (defined as in encoder-decoder framework)
- Neural Reasoner is better at positional reasoning and path finding

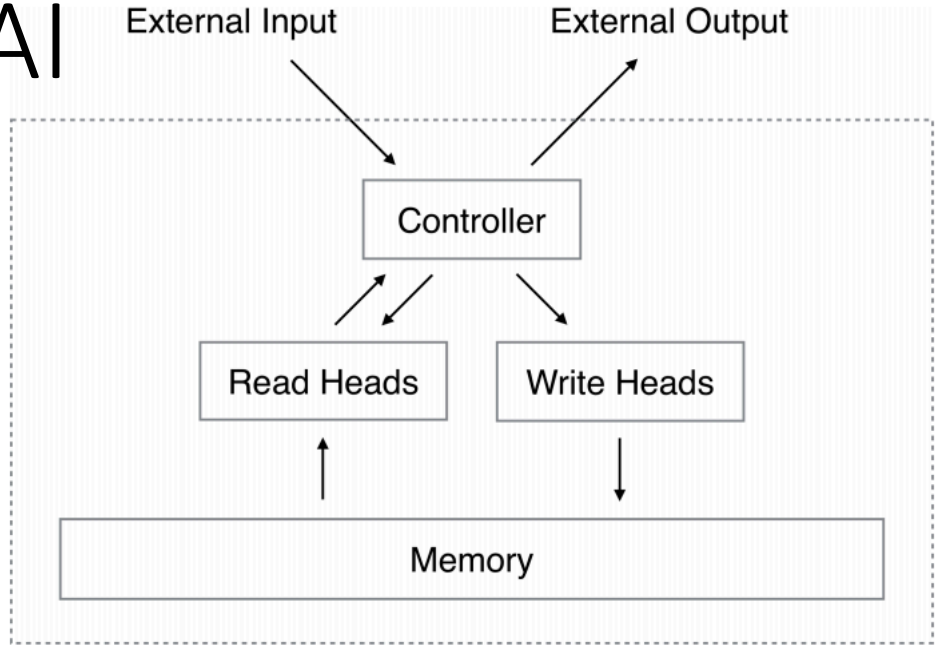
	Path Finding (1K)	Path Finding (10K)
No auxiliary task		
2-layer reasoning, 1-layer DNN	13.6%	52.2%
2-layer reasoning, 2-layer DNN	12.3%	54.2%
3-layer reasoning, 3-layer DNN	13.1%	51.7%
Auxiliary task: Original		
2-layer reasoning, 1-layer DNN	14.1%	57.0%
2-layer reasoning, 2-layer DNN	17.3%	87.0%
3-layer reasoning, 3-layer DNN	14.0%	98.4%
Auxiliary task: Abstract		
2-layer reasoning, 1-layer DNN	18.1%	55.8%
2-layer reasoning, 2-layer DNN	15.4%	87.8%
3-layer reasoning, 3-layer DNN	11.3%	98.6%



(a) reconstructing the original sentence b) producing an abstract form with variables

Neural Turing Machine – General AI

"Neural Turing Machine" Alex Graves et al (2014)

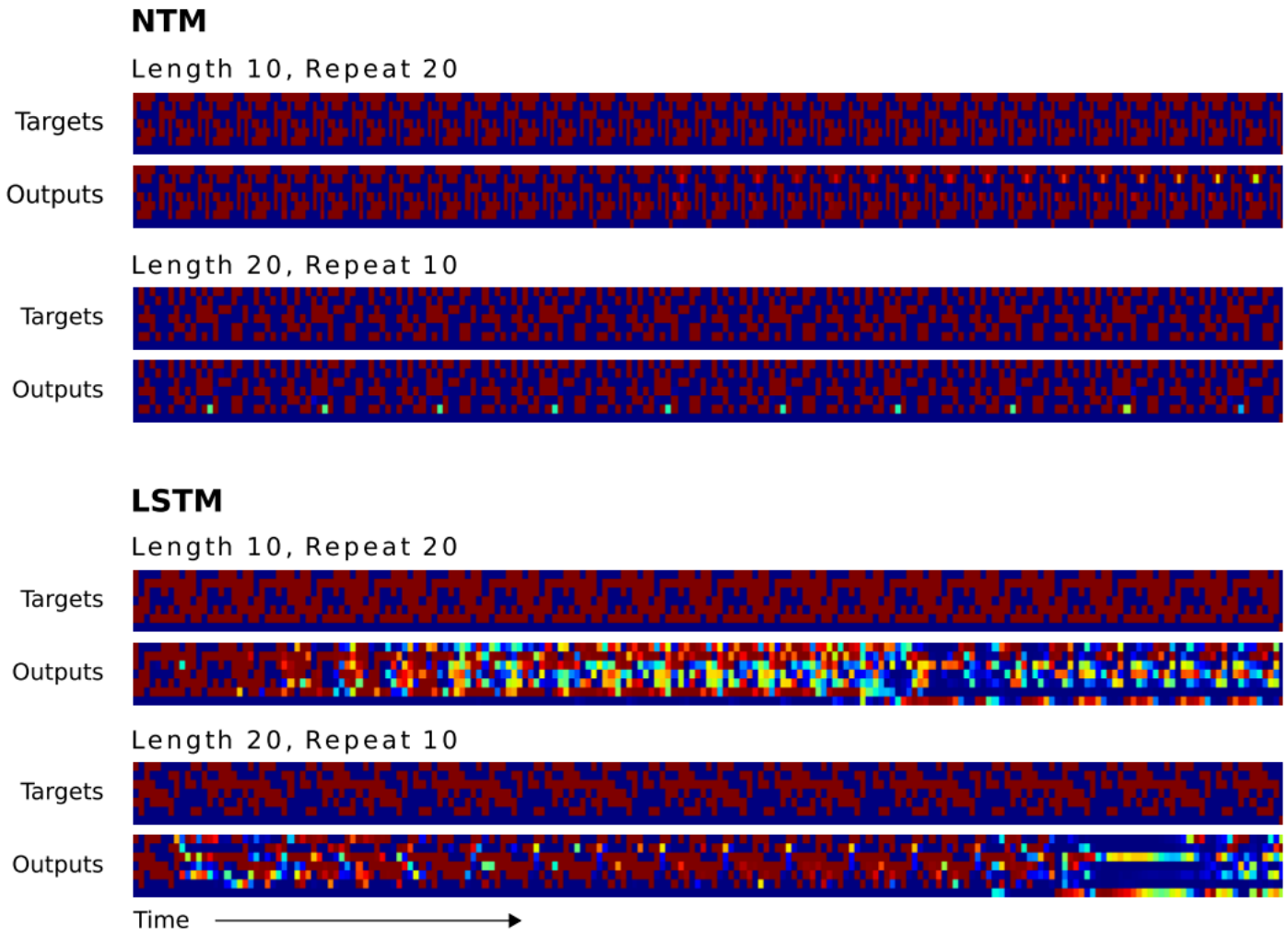


- Goal: infer operation of simple algorithms from series of input-output pairs
- Demonstrated ability to automatically learn to construct and iterate thru arrays, sort elements, copy sequences, and learn associations.
- Fully differentiable model with cross-entropy objective;
- **Controller NN** is given input/output pairs and reads/writes to a matrix of n **memory vectors**
  - Operations occur based on prob. distributions over the rows based on addressing system
  - **Location based addressing** (for generalization and intermediate calculations)
  - **Content based addressing** (if content in row is similar to a given key)
- Accomplished NL task of variable binding and recursive processing of variable length structures.
  - Gate determines whether to use location or content address
  - Grant ability to shift location to allow for iteration and random access
- NTM is more adept at generalizing training time tasks than LSTM, and with fewer parameters

1) content system can choose a weight without modificatic
2) weight from the content addressing system and be chos
allow for accessing a contiguous block of data and access a
that block. 3) a weighting from the previous time step can
input from the content system, allowing for iteration throu
addresses

Neural Turing Machine

"Neural Turing Machine" Alex Graves et al (2014)



NTM and LSTM Generalization for the Repeat Copy Task. NTM generalizes almost perfectly (target and output colors match) to longer sequences than seen during training.

Training task was to copy a length 10 sequence of 8 bit vectors 10 times. The test task was to repeat that sequence 20 times, or to copy a length 20 sequence 10 times – slightly different.

When the number of repeats is increased NTM is able to continue duplicating the input sequence fairly accurately; but it is unable to predict when the sequence will end, emitting the end marker after the end of every repetition beyond the eleventh.

LSTM struggles with both increased length and number, rapidly diverging from the input sequence in both cases

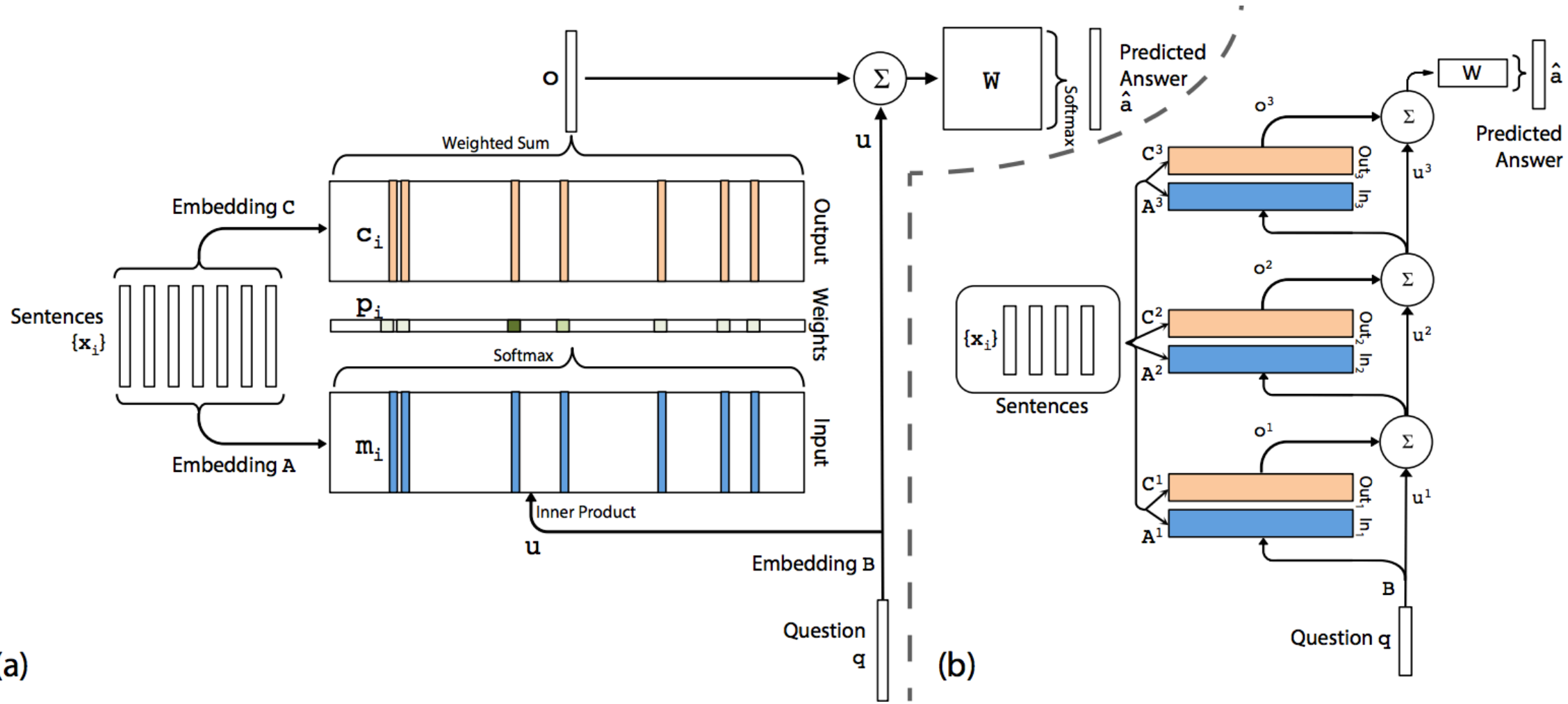
Attention Models for Fact Ranking

"End-to-end memory networks" Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015)

- End-to-end version of Weston's original MemNN architecture (no intermediate supervision)
- Vector representations of facts and the query embedded into same space are then processed via multiple hops. There are two different representations of facts: Input and Output.
  - The **internal** representation exists gauge the probability the fact is relevant to a query.
  - **Output** fact representations found by a different embedding used to compute the response vector to the query. The output vector is a combination of output representations weighted by probabilities computed from the internal representations.
- The final response vector is the combination of the fact-output vector and the query representation.
- **TASK 1: QA - BaBi dataset** each word of each supporting sentence encoded one-hot, same for query and answer. Bag of words sentence representations also used
- **TASK 2: Language Modeling -** predict the next word. Previous N words are embedded as separate "facts" in the memory array. No question exists, constant vector. output softmax predicts next word in vocab, cross entropy loss used again

Attention Models for Fact Ranking

"End-to-end memory networks" Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015)



(a): A single layer version of our model. (b): A three layer version of the model. In practice, we can constrain several of the embedding matrices to be the same – weight sharing to reduce parameters

Comparison of Strongly Supervised MemNN vs End-to-End MemNN

"End-to-end memory networks" Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015)

	Baseline			MemN2N								
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	PE LS	PE LS RN	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
1: 1 supporting fact	0.0	50.0	0.1	0.6	0.1	0.2	0.0	0.8	0.0	0.1	0.0	0.1
2: 2 supporting facts	0.0	80.0	42.8	17.6	21.6	12.8	8.3	62.0	15.6	14.0	11.4	18.8
3: 3 supporting facts	0.0	80.0	76.4	71.0	64.2	58.8	40.3	76.9	31.6	33.1	21.9	31.7
4: 2 argument relations	0.0	39.0	40.3	32.0	3.8	11.6	2.8	22.8	2.2	5.7	13.4	17.5
5: 3 argument relations	2.0	30.0	16.3	18.3	14.1	15.7	13.1	11.0	13.4	14.8	14.4	12.9
6: yes/no questions	0.0	52.0	51.0	8.7	7.9	8.7	7.6	7.2	2.3	3.3	2.8	2.0
7: counting	15.0	51.0	36.1	23.5	21.6	20.3	17.3	15.9	25.4	17.9	18.3	10.1
8: lists/sets	9.0	55.0	37.8	11.4	12.6	12.7	10.0	13.2	11.7	10.1	9.3	6.1
9: simple negation	0.0	36.0	35.9	21.1	23.3	17.0	13.2	5.1	2.0	3.1	1.9	1.5
10: indefinite knowledge	2.0	56.0	68.7	22.8	17.4	18.6	15.1	10.6	5.0	6.6	6.5	2.6
11: basic coreference	0.0	38.0	30.0	4.1	4.3	0.0	0.9	8.4	1.2	0.9	0.3	3.3
12: conjunction	0.0	26.0	10.1	0.3	0.3	0.1	0.2	0.4	0.0	0.3	0.1	0.0
13: compound coreference	0.0	6.0	19.7	10.5	9.9	0.3	0.4	6.3	0.2	1.4	0.2	0.5
14: time reasoning	1.0	73.0	18.3	1.3	1.8	2.0	1.7	36.9	8.1	8.2	6.9	2.0
15: basic deduction	0.0	79.0	64.8	24.3	0.0	0.0	0.0	46.4	0.5	0.0	0.0	1.8
16: basic induction	0.0	77.0	50.5	52.0	52.1	1.6	1.3	47.4	51.3	3.5	2.7	51.0
17: positional reasoning	35.0	49.0	50.9	45.4	50.1	49.0	51.0	44.4	41.2	44.5	40.4	42.6
18: size reasoning	5.0	48.0	51.3	48.1	13.6	10.1	11.1	9.6	10.3	9.2	9.4	9.2
19: path finding	64.0	92.0	100.0	89.7	87.4	85.6	82.8	90.7	89.9	90.2	88.0	90.6
20: agent's motivation	0.0	9.0	3.6	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2
Mean error (%)	6.7	51.3	40.2	25.1	20.3	16.3	13.9	25.8	15.6	13.3	12.4	15.2
Failed tasks (err. $>$ 5%)	4	20	18	15	13	12	11	17	11	11	11	10
On 10k training data												
Mean error (%)	3.2	36.4	39.2	15.4	9.4	7.2	6.6	24.5	10.9	7.9	7.5	11.0
Failed tasks (err. $>$ 5%)	2	16	17	9	6	4	4	16	7	6	6	6

Test error rates (%) on 1k training examples of BaBI.

Comparison of Strongly Supervised MemNN vs End-to-End MemNN

"End-to-end memory networks" Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015)

	Baseline			MemN2N									
Task	Strongly Supervised MemNN	LSTM	MemNN WSH	BoW	PE	PE LS	PE LS RN	PE LS LW RN*	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
1: 1 supporting fact	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2: 2 supporting facts	0.0	81.9	39.6	0.6	0.4	0.5	0.3	0.3	62.0	1.3	2.3	1.0	0.8
3: 3 supporting facts	0.0	83.1	79.5	17.8	12.6	15.0	9.3	2.1	80.0	15.8	14.0	6.8	18.3
4: 2 argument relations	0.0	0.2	36.6	31.8	0.0	0.0	0.0	0.0	21.4	0.0	0.0	0.0	0.0
5: 3 argument relations	0.3	1.2	21.1	14.2	0.8	0.6	0.8	0.8	8.7	7.2	7.5	6.1	0.8
6: yes/no questions	0.0	51.8	49.9	0.1	0.2	0.1	0.0	0.1	6.1	0.7	0.2	0.1	0.1
7: counting	3.3	24.9	35.1	10.7	5.7	3.2	3.7	2.0	14.8	10.5	6.1	6.6	8.4
8: lists/sets	1.0	34.1	42.7	1.4	2.4	2.2	0.8	0.9	8.9	4.7	4.0	2.7	1.4
9: simple negation	0.0	20.2	36.4	1.8	1.3	2.0	0.8	0.3	3.7	0.4	0.0	0.0	0.2
10: indefinite knowledge	0.0	30.1	76.0	1.9	1.7	3.3	2.4	0.0	10.3	0.6	0.4	0.5	0.0
11: basic coreference	0.0	10.3	25.3	0.0	0.0	0.0	0.0	0.1	8.3	0.0	0.0	0.0	0.4
12: conjunction	0.0	23.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
13: compound coreference	0.0	6.1	12.3	0.0	0.1	0.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0
14: time reasoning	0.0	81.0	8.7	0.0	0.2	0.0	0.0	0.1	30.9	0.2	0.2	0.0	1.7
15: basic deduction	0.0	78.7	68.8	12.5	0.0	0.0	0.0	0.0	42.6	0.0	0.0	0.2	0.0
16: basic induction	0.0	51.9	50.9	50.9	48.6	0.1	0.4	51.8	47.3	46.4	0.4	0.2	49.2
17: positional reasoning	24.6	50.1	51.1	47.4	40.3	41.1	40.7	18,6	40.0	39.7	41.7	41.8	40.0
18: size reasoning	2.1	6.8	45.8	41.3	7.4	8.6	6.7	5.3	9.2	10.1	8.6	8.0	8.4
19: path finding	31.9	90.3	100.0	75.4	66.6	66.7	66.5	2.3	91.0	80.8	73.3	75.7	89.5
20: agent's motivation	0.0	2.1	4.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean error (%)	3.2	36.4	39.2	15.4	9.4	7.2	6.6	4.2	24.5	10.9	7.9	7.5	11.0
Failed tasks (err. $>$ 5%)	2	16	17	9	6	4	4	3	16	7	6	6	6

Test error rates (%) on 10k training examples of BaBI.

Comparison of Strongly Supervised MemNN vs End-to-End MemNN

"End-to-end memory networks" Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015)

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom Prediction: bathroom				

Story (2: 2 supporting facts)	Support	Hop 1	Hop 2	Hop 3
John dropped the milk.		0.06	0.00	0.00
John took the milk there.	yes	0.88	1.00	0.00
Sandra went back to the bathroom.		0.00	0.00	0.00
John moved to the hallway.	yes	0.00	0.00	1.00
Mary went back to the bedroom.		0.00	0.00	0.00
Where is the milk? Answer: hallway Prediction: hallway				

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

Story (18: size reasoning)	Support	Hop 1	Hop 2	Hop 3
The suitcase is bigger than the chest.	yes	0.00	0.88	0.00
The box is bigger than the chocolate.		0.04	0.05	0.10
The chest is bigger than the chocolate.	yes	0.17	0.07	0.90
The chest fits inside the container.		0.00	0.00	0.00
The chest fits inside the box.		0.00	0.00	0.00
Does the suitcase fit in the chocolate? Answer: no Prediction: no				

Example predictions on the QA tasks of [22]. We show the labeled supporting facts (support) from the dataset which MemN2N does not use during training, and the probabilities p of each hop used by the model during inference. MemN2N successfully learns to focus on the correct supporting sentences.

Problem: Knowledge Representation Currently Insufficient for Complex Reasoning

How to represent entities and the relationships between them in a vector space
Smolensky: tensor products
Li Deng: embeddings

JOHNS HOPKINS
UNIVERSITY

18 September 2016

Current Vector-valued Knowledge Base Technology

Preserve graph structure; but find a representation more amenable to computation

JOHNS HOPKINS
UNIVERSITY

Tensor Products

"**Reasoning in Vector Space: An Exploratory Study of Question Answering**" Moontae Lee, Wen-tau Yih, Li Deng, Paul Smolensky

Illuminate what knowledge is captured in each representation. **Key: reason with transitivity-like rules**

Container-Containee Relationships solve a lot of the BaBi questions:

- represent distinct entities as d-dim unit vectors
- Bind containee entity to container entity via outer product: (containee)(container)^T

#	Statements/Questions	Relational Translations/Answers	Encodings/Clues	
1	Mary went to the kitchen.	*Mary belongs to the kitchen (from nowhere).*	mk^T	$m(k \circ n)^T$
3	Mary got the football there.	*The football belongs to Mary.*	fm^T	fm^T
4	Mary travelled to the garden.	*Mary belongs to the garden (from the kitchen).*	mg^T	$m(g \circ k)^T$
5	Where is the football?	garden	3, 4	
9	Mary dropped the football.	*The football belongs to where Mary belongs to.*	fg^T	fg^T
10	Mary journeyed to the kitchen.	*Mary belongs to the kitchen (from the garden).*	mk^T	$m(k \circ g)^T$
11	Where is the football?	garden	9, 4	

Sample containee-belongs-to-container translations and corresponding encodings about Mary. Symbols in encodings are all d-dimensional vectors for actors (mary), objects (football), and locations(nowhere, kitchen, garden). Translations and encodings for Category 3 are also specified with the parentheses and circle operation, respectively

$$(fm^T) \cdot (mg^T) = f(m^T \cdot m)g^T = fg^T \qquad (\because m^T m = ||m||_2^2 = 1)$$

Tensor Products

"**Reasoning in Vector Space: An Exploratory Study of Question Answering**" Moontae Lee, Wen-tau Yih, Li Deng, Paul Smolensky

Temporal Encoding:

- When an entity changes location from *prev* (p) to *next* (n), bind this transition into a d-dim vector using a *d x 2d* temporal encoding matrix, U

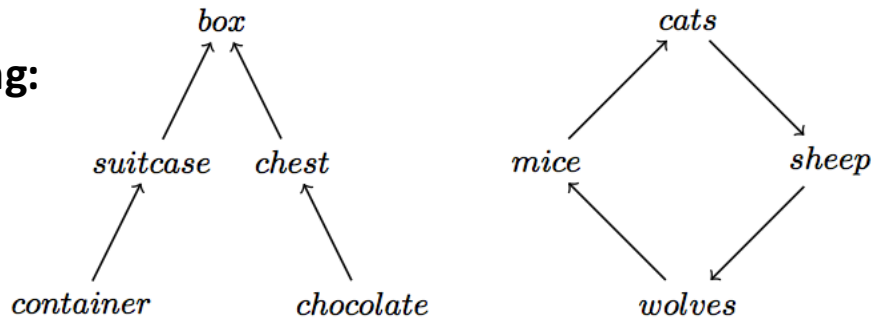
$$n \circ p = U \begin{bmatrix} n \\ p \end{bmatrix} \in \mathbb{R}^d$$

Mary belongs to the garden (from the kitchen): $m(g \circ k)^T$.

#	Statements/Questions	Relational Translations/Answers	Encodings/Clues	
1	Mary went to the kitchen.	*Mary belongs to the kitchen (from nowhere).*	mk^T	$m(k \circ n)^T$
3	Mary got the football there.	*The football belongs to Mary.*	fm^T	fm^T
4	Mary travelled to the garden.	*Mary belongs to the garden (from the kitchen).*	mg^T	$m(g \circ k)^T$
5	Where is the football?	garden	3, 4	
9	Mary dropped the football.	*The football belongs to where Mary belongs to.*	fg^T	fg^T
10	Mary journeyed to the kitchen.	*Mary belongs to the kitchen (from the garden).*	mk^T	$m(k \circ g)^T$
11	Where is the football?	garden	9, 4	

Sample containee-belongs-to-container translations and corresponding encodings about Mary. Symbols in encodings are all d-dimensional vectors for actors (mary), objects (football), and locations(nowhere, kitchen, garden). Translations and encodings for Category 3 are also specified with the parentheses and circle operation, respectively

All deduction tasks can be solved with the container-containee encoding:



Tensor Products

"**Reasoning in Vector Space: An Exploratory Study of Question Answering**" Moontae Lee, Wen-tau Yih, Li Deng, Paul Smolensky

Ownership Transfer – analogous to location change

- When an entity changes owners from *prev* (p) to *next* (n), bind this transition into a d-dim vector using a *d x 2d* temporal encoding matrix, V

Conjunctions – multiple actors

- Conjoin two objects by another bilinear binding operation $\star : R^d \times R^d \rightarrow R^d$, and unbind similarly via the pseudo-inverse of the corresponding matrix

Indefinite Knowledge:

- If two locations unbound from the target actor are identical, we output a yes/no definite answer, whereas two different locations imply the indefinite answer 'maybe' if one of the unbound locations matches the queried location

#	Statements/Questions	Relational Translations/Answers	Encodings/Clues
1	Jeff took the milk there.	*The milk belongs to Jeff (from None).*	$m(j * n)^T$
2	Jeff gave the milk to Bill.	*The milk belongs to Bill (from Jeff).*	$m(b * j)^T$
3	Who did Jeff give the milk to?	Bill	2
4	Daniel travelled to the office.	*Daniel belongs to the office.*	do^T
5	Daniel journeyed to the hallway.	*Daniel belongs to the hallway.*	dh^T
6	Who received the milk?	Bill	2
7	Bill went to the kitchen.	*Bill belongs to the kitchen.*	bk^T
8	Fred grabbed the apple there.	*The apple belongs to Fred (from none).*	$a(f * n)^T$
9	What did Jeff give to Bill?	milk	2

Category 12: Conjunction
01: Daniel and Sandra went back to the kitchen.
02: Daniel and John went back to the hallway.
03: Where is Daniel? hallway 2
04: Daniel and John moved to the bathroom.
05: Sandra and Mary travelled to the office.
06: Where is Daniel? bathroom 4

Category 10: Indefinite Knowledge
01: Julie travelled to the kitchen.
02: Bill is either in the school or the office.
03: Is Bill in the office? maybe 2
04: Bill went back to the bedroom.
05: Bill travelled to the kitchen.
06: Is Bill in the kitchen? yes 5

Tensor Products

"**Reasoning in Vector Space: An Exploratory Study of Question Answering**" Moontae Lee, Wen-tau Yih, Li Deng, Paul Smolensky

Induction: Categories 16 and 20 - but this is just the container-containee relationship applied in *reverse*

> Similarly in Category 20, there exists precisely one statement which describes a property of an actor (e.g., *"Sumit is bored."* $= bs^T$). Then a statement describes the actor's relocation (e.g., *"Sumit journeyed to the garden."* $= sg^T$), yielding an inductive conclusion by matrix multiplication: *"Being boring makes people go to the garden."* $= (bs^T) \cdot (sg^T) = bg^T$. The inductive reasoning also generalizes to other actions (e.g., the reason for later activity, *"Sumit grabbed the football."* $= sf^T$, is also being bored, because $(bs^T) \cdot (sf^T) = bf^T$).

Path finding: Deferential computation

- If the location of both entities in a direction are not known, store them in a queue. If the location of one is known relative to an unknown location, multiply by direction matrix

#	Statements/Questions	Translations/Answers/Clues	Encodings	Seq
1	The bedroom is south of the hallway.	*Decides b given the initial h.*	$b = Sh$	(1)
2	The βathroom is east of the office.	*Defer until we know either o or β.*	$\beta = Eo$	(3)
3	The kitchen is west of the garden.	*Defer until we know either g or k.*	$k = Wg$	(5)
4	The garden is south of the office.	*Defer until we know either o or g.*	$g = So$	(4)
5	The office is south of the bedroom.	*Decides o given b.*	$o = Sb$	(2)
6	How do you go from the garden to the bedroom?	n,n 4, 5	$b = Xg$	(6)

Sample multi-relational translations and corresponding encodings from Category 19. Symbols in encodings are either d-dimensional object vectors (hallway, bedroom, office, βathroom, garden, kitchen) or d × d directional matrices (South, East, West, North). The last column shows the sequence of actual running order

Tensor Products

"**Reasoning in Vector Space: An Exploratory Study of Question Answering**" Moontae Lee, Wen-tau Yih, Li Deng, Paul Smolensky

Tensor product models are superior, but require a new tensor operation for every "type" of question, which requires nontrivial human ingenuity

Type	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Training	100%	100%	100%	100%	99.8%	100%	100%	100%	100%	100%
Test	100%	100%	100%	100%	99.8%	100%	100%	100%	100%	100%
Type	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**
Training	100%	100%	100%	100%	100%	99.4%	100%	100%	100%	100%
Test	100%	100%	100%	100%	100%	99.5%	100%	100%	100%	100%

Accuracies on training and test data on **Tensor Product** models. They achieve near-perfect accuracy in almost every category including positional reasoning and path finding.

Type	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Accuracy	100%	100%	100%	100%	99.3%	100%	96.9%	96.5%	100%	99%
Model	MNN	MNN	MNN	MNN	DMN	MNN	DMN	DMN	DMN	SSVM
Type	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**
Accuracy	100%	100%	100%	100%	100%	100%	72%	95%	36%	100%
Model	MNN	MNN	MNN	DMN	MNN	MNN	Multitask	MNN	MNN	MNN

Best accuracies achieved by competing models. MNN = Strongly-Supervised MemNN trained with the clue numbers, DMN = Dynamic MemNN, SSVM =Structured SVM with the coreference resolution and SRL features. Multitask indicates multitask training.

Entity and Relationship Embeddings

"Embeddings Entities and Relations for Learning and Inference in Knowledge Bases" Wen-tau Yih, Xiaodong He, Jianfeng Gao, Li Deng

Learn vector representations of (subject, relationship, object) triples for KB. Model Relations are linear or bilinear maps. Purpose: **deduce new facts** and complete KB, **support complex reasoning in soft vector space**, provide explanations for inference results

- Link Prediction: entity ranking task - predict correctness of unseen triples
- Horn Rule extraction: mine new relationships that complete a closed path in the KB
  - B(a, b) and C(c, d) and D(e, f) => H(a, f) where H is a new relationship that exists between 'a' and 'f'
- Objective: min margin-based ranking objective, which encourages the scores of positive relationships (triplets) to be higher than the scores of any negative relationships (triplets)

$$g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) = \mathbf{A}_r^T \begin{pmatrix} \mathbf{y}_{e_1} \\ \mathbf{y}_{e_2} \end{pmatrix} \quad \text{and} \quad g_r^b(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) = \mathbf{y}_{e_1}^T \mathbf{B}_r \mathbf{y}_{e_2}$$

Two different relationship scoring models, where A and B are the linear and bilinear relatinoship operators, respectively.

$$\mathbf{y}_{e_1} = f\left(\mathbf{W}\mathbf{x}_{e_1}\right), \quad \mathbf{y}_{e_2} = f\left(\mathbf{W}\mathbf{x}_{e_2}\right)$$

Representation of two entities where W is some parameter matrix

Models	$\mathbf{B}_r$	$\mathbf{A}_r^T$	Scoring Function
Distance (Bordes et al., 2011)	-	$\left(\mathbf{Q}_{r_1}^T \quad -\mathbf{Q}_{r_2}^T\right)$	$-\|\|g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2})\|\|_1$
Single Layer (Socher et al., 2013)	-	$\left(\mathbf{Q}_{r1}^T \quad \mathbf{Q}_{r2}^T\right)$	$\mathbf{u}_r^T \tanh(g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}))$
TransE (Bordes et al., 2013b)	$\mathbf{I}$	$\left(\mathbf{V}_r^T \quad -\mathbf{V}_r^T\right)$	$-(2g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) - 2g_r^b(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) + \|\|\mathbf{V}_r\|\|_2^2)$
NTN (Socher et al., 2013)	$\mathbf{T}_r$	$\left(\mathbf{Q}_{r1}^T \quad \mathbf{Q}_{r2}^T\right)$	$\mathbf{u}_r^T \tanh\left(g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) + g_r^b(\mathbf{y}_{e_1}, \mathbf{y}_{e_2})\right)$

Comparison of different scoring functions based on the relationship models

Entity and Relationship Embeddings

"Embeddings Entities and Relations for Learning and Inference in Knowledge Bases" Wen-tau Yih, Xiaodong He, Jianfeng Gao, Li Deng

Examples of length-2 rules extracted by EMBEDRULE with embeddings learned from DISTMULT-tanh-EV-init:

$$AwardInCeremany(a,b) \wedge CeremanyEventType(b,c) \implies AwardInEventType(a,c)$$
$$AtheletePlayInTeam(a,b) \wedge TeamPlaySport(b,c) \implies AtheletePlaySport(a,c)$$
$$TVProgramInTVNetwork(a,b) \wedge TVNetworkServiceLanguage(b,c) \implies TVProgramLanguage(a,c)$$
$$LocationInState(a,b) \wedge StateInCountry(b,c) \implies LocationInCountry(a,c)$$
$$BornInLocation(a,b) \wedge LocationInCountry(b,c) \implies Nationality(a,c)$$

Examples of length-3 rules extracted by EMBEDRULE with embeddings learned from DISTMULT-tanh-EV-init:

$$SportPlayByTeam(a,b) \wedge TeamInClub(b,c) \wedge ClubHasPlayer(c,d) \implies SportPlayByAthelete(a,d)$$
$$MusicTrackPerformer(a,b) \wedge PeerInfluence(b,c) \wedge PerformRole(c,d) \implies MusicTrackRole(a,d)$$
$$FilmHasActor(a,b) \wedge CelebrityFriendship(b,c) \wedge PersonLanguage(c,d) \implies FilmLanguage(a,d)$$

Suggestions for Future Research