Memory Networks for Language Understanding

Jason Weston

Facebook AI Research
Intelligent Conversational Agents

What can I help you with?

“Play a good song”

Sorry, I couldn’t find ‘a good song’ in your music.
End-to-End Dialog Agents

While it is possible to build useful dialog agents as a set of separate black boxes with joining logic (Google Now, Cortana, Siri, .. ?) we believe a true dialog agent should:

- Be able to combine all its knowledge to fulfill complex tasks.

- Handle long open-ended conversations involving effectively tracking many latent variables.

- Be able to learn (new tasks) via conversation.

*Our bet:* Machine Learning End-to-End systems is the way forward in the long-run.
Memory Networks

- Class of models that combine large memory with learning component that can read and write to it.
- Incorporates reasoning with attention over memory (RAM).
- Most ML has limited memory which is more-or-less all that’s needed for “low level” tasks e.g. object detection.

Our motivation: long-term memory is required to read a story and then e.g. answer questions about it.

Similarly, it’s also required for dialog: to remember previous dialog (short- and long-term), and respond.

1. We first test this on the toy (bAbI) tasks.
2. Any interesting model has to be good on real data as well.
## Memory Networks

| Long-Term Memories \( h_i \) | Shaolin Soccer directed by Stephen Chow  
| Shaolin Soccer written by Stephen Chow  
| Shaolin Soccer starred actors Stephen Chow  
| Shaolin Soccer release year 2001  
| Shaolin Soccer has genre comedy  
| Shaolin Soccer has tags martial arts, kung fu soccer, stephen chow  
| Kung Fu Hustle directed by Stephen Chow  
| Kung Fu Hustle written by Stephen Chow  
| Kung Fu Hustle starred actors Stephen Chow  
| Kung Fu Hustle has genre comedy action  
| Kung Fu Hustle has imdb votes famous  
| Kung Fu Hustle has tags comedy, action, martial arts, kung fu, china, soccer, hong kong, stephen chow  
| The God of Cookery directed by Stephen Chow  
| The God of Cookery written by Stephen Chow  
| The God of Cookery starred actors Stephen Chow  
| The God of Cookery has tags hong kong Stephen Chow  
| From Beijing with Love directed by Stephen Chow  
| From Beijing with Love written by Stephen Chow  
| From Beijing with Love starred actors Stephen Chow, Anita Yuen  
| ... <and more> ... |

| Short-Term Memories \( c_j^* \) |  
| 1) I’m looking a fun comedy to watch tonight, any ideas?  
| 2) Have you seen Shaolin Soccer? That was zany and great.. really funny but in a whacky way.  
| 3) Yes! Shaolin Soccer and Kung Fu Hustle are so good I really need to find some more Stephen Chow films I feel like there is more awesomeness out there that I haven’t discovered yet ...  
| Output \( y \) | 4) God of Cookery is pretty great, one of his mid 90’s hong kong martial art comedies. |
Evaluating End-To-End Learners

- **Long Term goal:** A learner can be trained (from scratch?) to understand and use language.

- **Our main interest:** uncover the learning algorithms able to do so.

- **Inspired by “A Roadmap towards Machine Intelligence”** (Mikolov, Joulin, Baroni 2015) we advocate a set of tasks to train & evaluate on:
  - **Classic Language Modeling** (Penn TreeBank, Text8)
  - **Story understanding** (Children’s Book Test, News articles)
  - **Open Question Answering** (WebQuestions, WikiQA)
  - **Goal-Oriented Dialog and Chit-Chat** (Movie Dialog, Ubuntu)
What is a Memory Network?

Original paper description of class of models

MemNNs have four component networks (which may or may not have shared parameters):

- **I**: (input feature map) convert incoming data to the internal feature representation.
- **G**: (generalization) update memories given new input.
- **O**: produce new output (in feature representation space) given the memories.
- **R**: (response) convert output O into response seen by the outside world.
Some Memory Network-related Publications

Memory Network Models
implemented models..

Memory Module
Supervision (direct or reward-based)
Output

Controller module

Memory Module
$m$
read
addressing

$m$
read
addressing
$\{\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_N\}$

Input
Internal state Vector (initially: query)

[Figure by Saina Sukhbaatar]
Variants of the class...

Some options and extensions:

- **Representation of inputs and memories could use all kinds of encodings:** bag of words, RNN style reading at word or character level, etc.

- **Different possibilities for output module:** e.g. multi-class classifier or uses an RNN to output sentences.

- **If the memory is huge** (e.g. Wikipedia) we need to organize the memories. Solution: hash the memories to store in buckets (topics). Then, memory addressing and reading doesn’t operate on all memories.

- **If the memory is full**, there could be a way of removing one it thinks is most useless; *i.e.* it `forgets’’ somehow. That would require a scoring function of the utility of each memory.
Task (1) Factoid QA with Single Supporting Fact (‘‘where is actor’’)

(Very Simple) Toy reading comprehension task:

John was in the bedroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A:kitchen
A harder (toy) task is to answer questions where two supporting statements have to be chained to answer the question:

John is in the playground.
Bob is in the office.
John picked up the football.
Bob went to the kitchen.
Where is the football? A: playground
Where was Bob before the kitchen? A: office
(2) Factoid QA with Two Supporting Facts ("where is actor+object")

A harder (toy) task is to answer questions where two supporting statements have to be chained to answer the question:

John is in the playground.  
Bob is in the office.  
John picked up the football.  
Bob went to the kitchen.  
Where is the football?  A: playground
Where was Bob before the kitchen?  A: office

To answer the first question Where is the football? both John picked up the football and John is in the playground are supporting facts.
Memory Network Models

implemented models

---

Supervision
(direct or
reward-based)

Output

Controller module

\[ \vec{u}_2 \]

\[ \vec{u}_1 \]

Memory Module

\[ m \]

\( \{ \tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_N \} \)

Input

Memory vectors

Internal state Vector (initially: query)

[Figure by Saina Sukhbaatar]
The First MemNN Implementation

- **I** (input): converts to bag-of-word-embeddings $x$.
- **G** (generalization): stores $x$ in next available slot $m_N$.
- **O** (output): Loops over all memories $k=1$ or 2 times:
  - 1$^{\text{st}}$ loop max: finds best match $m_i$ with $x$.
  - 2$^{\text{nd}}$ loop max: finds best match $m_J$ with $(x, m_i)$.
  - The output $o$ is represented with $(x, m_i, m_J)$.
- **R** (response): ranks all words in the dictionary given $o$ and returns best single word. *(OR: use a full RNN here)*
Matching function

- For a given Q, we want a good match to the relevant memory slot(s) containing the answer, e.g.:

Match(Where is the football ?, John picked up the football)

- We use a $q^T U^T U d$ embedding model with word embedding features:
  - **LHS features**: Q:Where Q:is Q:the Q:football Q:?  
  - **RHS features**: D:John D:picked D:up D:the D:football  
    QDMatch:the QDMatch:football  
    (QDMatch:football is a feature to say there’s a Q&A word match, which can help.)

The parameters U are trained with a margin ranking loss: supporting facts should score higher than non-supporting facts.
Matching function: $2^{nd}$ hop

- On the $2^{nd}$ hop we match question & $1^{st}$ hop to new fact:

  Match( [Where is the football ?, John picked up the football], John is in the playground)

- We use the same $q^T U^T U_d$ embedding model:
  - **LHS features**: Q:Where Q:is Q:the Q:football Q:? Q2:John Q2:picked Q2:up Q2:the Q2:football
  - **RHS features**: D:John D:is D:in D:the D:playground QDMatch:the QDMatch:is .. Q2DMatch:John
Objective function

Minimize:

\[
\sum_{\bar{f} \neq m_{o1}} \max(0, \gamma - s_O(x, m_{o1}) + s_O(x, \bar{f})) + \\
\sum_{\bar{f}' \neq m_{o2}} \max(0, \gamma - s_O([x, m_{o1}], m_{o2}) + s_O([x, m_{o1}], \bar{f}'])) + \\
\sum_{\bar{r} \neq r} \max(0, \gamma - s_R([x, m_{o1}, m_{o2}], r) + s_R([x, m_{o1}, m_{o2}], \bar{r}']))
\]

Where:
- $S_O$ is the matching function for the Output component.
- $S_R$ is the matching function for the Response component.
- $x$ is the input question.
- $m_{o1}$ is the first true supporting memory (fact).
- $m_{o2}$ is the first second supporting memory (fact).
- $r$ is the response.

True facts and responses $m_{o1}$, $m_{o2}$ and $r$ should have higher scores than all other facts and responses by a given margin.
Comparing triples

- We also need time information for the bAbI tasks. We tried adding absolute time as a feature: it works, but the following idea can be better:

- Seems to work better if we compare triples:

- $\text{Match}(Q,D,D')$ returns $< 0$ if $D$ is better than $D'$
  returns $> 0$ if $D'$ is better than $D$

We can loop through memories, keep best $m_i$ at each step.

Now the features include relative time features:

$L.H.S$: same as before

$R.H.S$: $\text{features}(D)$ $D_{\text{before}Q}$: 0-or-1
  $\text{features}(D')$ $D'_{\text{before}Q}$: 0-or-1 $D_{\text{before}D'}$: 0-or-1
Comparing triples: Objective and Inference

\[
\sum_{\bar{f} \neq m_{o_1}} \max(0, \gamma - s_{O_t}(x, m_{o_1}, \bar{f})) + \sum_{\bar{f} \neq m_{o_1}} \max(0, \gamma + s_{O_t}(x, \bar{f}, m_{o_1})) + \\
\sum_{\bar{f}' \neq m_{o_2}} \max(0, \gamma - s_{O_t}([x, m_{o_1}], m_{o_2}, \bar{f}')) + \sum_{\bar{f}' \neq m_{o_2}} \max(0, \gamma + s_{O_t}([x, m_{o_1}], \bar{f}', m_{o_2})) + \\
\sum_{\bar{r} \neq r} \max(0, \gamma - s_{R}([x, m_{o_1}, m_{o_2}], r) + s_{R}([x, m_{o_1}, m_{o_2}], \bar{r}]))
\]

Similar to before, except now for both \(m_{o_1}\) and \(m_{o_2}\) we need to have two terms considering them as the second or third argument to the \(S_{O_t}\) as they may appear on either side during inference:

**Algorithm 1** \(O_t\) replacement to arg max when using write time features

```
function \(O_t(q, m)\)
    \(t \leftarrow 1\)
    for \(i = 2, \ldots, N\) do
        if \(s_{O_t}(q, m_i, m_t) > 0\) then
            \(t \leftarrow i\)
        end if
    end for
    return \(t\)
end function
```
bAbI Experiment 1

- 10k sentences. (Actor: only ask questions about actors.)
- Difficulty: how many sentences in the past when entity mentioned.
- Fully supervised (supporting sentences are labeled).
- Compare RNN (no supervision) and MemNN hops $k = 1$ or $2$, & with/without time features.

<table>
<thead>
<tr>
<th>Method</th>
<th>Difficulty 1</th>
<th>Difficulty 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>actor</td>
<td>actor+object</td>
</tr>
<tr>
<td>RNN</td>
<td>0%</td>
<td>42%</td>
</tr>
<tr>
<td>MemNN $k = 1$</td>
<td>10%</td>
<td>81%</td>
</tr>
<tr>
<td>MemNN $k = 1$ (+time)</td>
<td>0%</td>
<td>27%</td>
</tr>
<tr>
<td>MemNN $k = 2$ (+time)</td>
<td>0%</td>
<td>0.05%</td>
</tr>
</tbody>
</table>

Difficulty 5 -- Max mem. sz. required: 65  Average mem. sz. required: 9
Example test story + predictions:

Antoine went to the kitchen. Antoine got the milk. Antoine travelled to the office. Antoine dropped the milk. Sumit picked up the football. Antoine went to the bathroom. Sumit moved to the kitchen.

- where is the milk now? A: office
- where is the football? A: kitchen
- where is Antoine? A: bathroom
- where is Sumit? A: kitchen
- where was Antoine before the bathroom? A: office
Joe went to the garden then Fred picked up the milk; Joe moved to the bathroom and Fred dropped the milk, and then Dan moved to the living room.

- Where is Dan? A: living room I believe
- Where is Joe? A: the bathroom

Fred moved to the bedroom and Joe went to the kitchen then Joe took the milk there and Dan journeyed to the bedroom; Joe discarded the milk.

- Where is the milk now? A: the milk is in the kitchen
- Where is Dan now? A: I think he is in the bedroom

Joe took the milk there, after that Mike travelled to the office, then Joe went to the living room, next Dan went back to the kitchen and Joe travelled to the office.

- Where is Joe now? A: I think Joe is in the office
Larger QA: Reverb Dataset in (Fader et al., 13)

- 14M statements, stored as (subject, relation, object) triples. Triples are REVERB extractions mined from ClueWeb09.

- Statements cover diverse topics:
  - (milne, authored, winnie-the-pooh)
  - (sheep, be-afraid-of, wolf), etc...

- Weakly labeled QA pairs and 35M paraphrased questions from WikiAnswers:
  - ``Who wrote the Winnie the Pooh books?``
  - ``Who is poohs creator?''
Results: QA on Reverb data from (Fader et al.)

• 14M statements stored in the memNN memory.
• $k=1$ loops MemNN, 128-dim embedding.
• R response simply outputs top scoring statement.
• Time features are not necessary, hence not used.
• We also tried adding bag of words (BoW) features.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Fader et al., 2013)</td>
<td>0.54</td>
</tr>
<tr>
<td>(Bordes et al., 2014)</td>
<td>0.73</td>
</tr>
<tr>
<td>MemNN</td>
<td>0.72</td>
</tr>
<tr>
<td>MemNN (with BoW features)</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Fast QA on Reverb data

Scoring all 14M candidates in the memory is slow.

We consider speedups using hashing in \( S \) and \( O \) as mentioned earlier:

- Hashing via words (essentially: inverted index)
- Hashing via k-means in embedding space (k=1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>Embedding</th>
<th>Embed+BoW</th>
<th>candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>MemNN (no hashing)</td>
<td>0.72</td>
<td>0.82</td>
<td>14M</td>
</tr>
<tr>
<td>MemNN (word hash)</td>
<td>0.63</td>
<td>0.68</td>
<td>13k (1000x)</td>
</tr>
<tr>
<td>MemNN (clust hash)</td>
<td>0.71</td>
<td>0.80</td>
<td>177k (80x)</td>
</tr>
</tbody>
</table>
A MemNN multitasked on bAbI data and Reverb QA data

The “story” told to the model after training:

Antoine went to the kitchen. Antoine picked up the milk. Antoine travelled to the office.

MemNN’s answers to some questions:

- Where is the milk?  A: office
- Where was Antoine before the office? A: kitchen
- Where does milk come from?  A: milk come from cow
- What is a cow a type of?  A: cow be female of cattle
- Where are cattle found?  A: cattle farm become widespread in brazil
- What does milk taste like? A: milk taste like milk
- What does milk go well with? A: milk go with coffee
Related Memory Models

(published before or ~same time as original paper)

- **RNNSearch (Bahdanau et al.)** for Machine Translation
  - Can be seen as a Memory Network where memory goes back only one sentence (writes embedding for each word).
  - At prediction time, reads memory and performs a soft max to find best alignment (most useful words). 1 hop only.

- **Generating Sequences With RNNs (Graves, ‘13)**
  - Also does alignment with previous sentence to generate handwriting (so RNN knows what letter it’s currently on).

- **Neural Turing Machines (Graves et al., 14)**
  - Has read and write operations over memory to perform tasks (e.g. copy, sort, associative recall).
  - 128 memory slots in experiments; content addressing computes a score for each slot ➔ slow for large memory?

- Earlier work by (Das ‘92), (Schmidhuber et al., 93), DISCERN (Miikkulainen, ‘90) and others...
Learning of Basic Algorithms using Reasoning, Attention, Memory (RAM) (e.g. addition, multiplication, sorting)  

Methods include adding stacks and addressable memory to RNNs:

- “Neural Net Architectures for Temporal Sequence Processing.” M. Mozer.
- “Neural Turing Machines” A. Graves, G. Wayne, I. Danihelka.
- “Inferring Algorithmic Patterns with Stack Augmented Recurrent Nets.” A. Joulin, T. Mikolov.
- “Learning to Transduce with Unbounded Memory” E. Grefenstette et al.
- “Neural Programmer-Interpreters” S. Reed, N. de Freitas.
- “The Neural GPU and the Neural RAM machine” I. Sutskever.
Classic NLP tasks for RAM

Classic Language Modeling:


Machine translation:

- “Sequence to Sequence Learning with Neural Networks” I. Sutskever, O. Vinyals, Q. Le.

Parsing:


Entailment:


Summarization:

- “A Neural Attention Model for Abstractive Sentence Summarization” A. M. Rush, S. Chopra, J. Weston.
Reasoning with synthetic language

- “A Roadmap towards Machine Intelligence” T. Mikolov, A. Joulin, M. Baroni.

Several new models that attempt to solve bAbI tasks:

- “End-To-End Memory Networks” S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus.
New NLP Datasets for RAM

Understanding news articles:


Understanding children’s books:


Conducting Dialog:


- “A Neural Network Approach to Context-Sensitive Generation of Conversational Responses” Sordoni et al.


- “Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems” J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, J. Weston.

General Question Answering:

- “Large-scale Simple Question Answering with Memory Networks” A. Bordes, N. Usunier, S. Chopra, J. Weston.
What was next for MemNNs?

• Make the language much harder: coreference, conjunctions, negations, etc. etc – *will it work?*

• MemNNs that reason with *more than* 2 supporting memories.

• End-to-end? (doesn’t need supporting facts)

• More useful applications on real datasets.

• Dialog: Ask questions? Say statements?

• *Do MemNN ideas extend to other ML tasks and model variants, e.g. visual QA, perform actions...?* [A: yes!].
bAbI tasks: what reasoning tasks would we like models to work on?

- We define 20 tasks (generated by the simulation) that we can test new models on. (See: http://fb.ai/babi)

- The idea is they are a bit like software tests: each task checks if an ML system has a certain skill.

- We would like each “skill” we check to be a natural task for humans w.r.t. text understanding & reasoning, humans should be able to get 100%.

Simulation commands

- go <place>
- get <object>
- get <object1> from <object2>
- put <object1> in/on <object2>
- give <object> to <person>
- drop <object>
- look
- inventory
- examine <object>

+ 2 commands for "gods" (superusers):
- create <object>
- set <obj1> <relation> <obj2>
Jason went to the kitchen.
Jason picked up the milk.
Jason travelled to the office.
Jason left the milk there.
Jason went to the bathroom.
Where is the milk now?  A: office
Where is Jason?  A: bathroom
Task (1) Factoid QA with Single Supporting Fact (‘‘where is actor’’)

Our first task consists of questions where a single supporting fact, previously given, provides the answer.

We test simplest case of this, by asking for the location of a person.

A small sample of the task is thus:

John is in the playground.
Bob is in the office.
Where is John? A: playground

We could use supporting facts for supervision at training time, but are not known at test time (we call this “strong supervision”). However weak supervision is much better!!
(2) Factoid QA with Two Supporting Facts (“where is actor+object”)

A harder task is to answer questions where two supporting statements have to be chained to answer the question:

John is in the playground.
Bob is in the office.
John picked up the football.
Bob went to the kitchen.
Where is the football? A: playground

To answer the question *Where is the football?* both *John picked up the football* and *John is in the playground* are supporting facts.
(3) Factoid QA with Three Supporting Facts

Similarly, one can make a task with 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

The first three statements are all required to answer this.
(4) Two Argument Relations: Subject vs. Object

To answer questions the ability to differentiate and recognize subjects and objects is crucial.

We consider the extreme case: sentences feature re-ordered words:

```
The office is north of the bedroom.
The bedroom is north of the bathroom.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom
```

Note that the two questions above have exactly the same words, but in a different order, and different answers.

So a bag-of-words will not work.
This task tests, in the simplest case possible (with a single supporting fact) the ability of a model to answer true/false type questions:

John is in the playground.
Daniel picks up the milk.
Is John in the classroom? A:no
Does Daniel have the milk? A:yes
(7) Counting

Tests ability to count sets:

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

(8) Lists/Sets

Tests ability to produce lists/sets:

Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
What is Daniel holding? A:milk,football
Basic Coreference (nearest referent)

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

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
(14) Time manipulation

- While our tasks so far have included time implicitly in the order of the statements, this task tests understanding the use of time expressions within the statements:

<table>
<thead>
<tr>
<th>In the afternoon Julie went to the park.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yesterday Julie was at school.</td>
</tr>
<tr>
<td>Julie went to the cinema this evening.</td>
</tr>
<tr>
<td>Where did Julie go after the park? A: cinema</td>
</tr>
</tbody>
</table>

**Much harder difficulty:** adapt a real time expression labeling dataset into a question answer format, e.g. Uzzaman et al., ‘12.
(15) Basic Deduction

- This task tests basic deduction via inheritance of properties:

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

Deduction should prove difficult for MemNNs because it effectively involves search, although our setup might be simple enough for it.
(17) Positional Reasoning

- This task tests spatial reasoning, one of many components of the classical SHRDLU system:

  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
(18) Reasoning about size

- This task requires reasoning about relative size of objects and is inspired by the commonsense reasoning examples in the Winograd schema challenge:

| The football fits in the suitcase. |
| The suitcase fits in the cupboard. |
| The box of chocolates is smaller than the football. |
| Will the box of chocolates fit in the suitcase? A:yes |

Tasks 3 (three supporting facts) and 6 (Yes/No) are prerequisites.
(19) Path Finding

In this task the goal is to find the path between locations:

The kitchen is north of the hallway.
The den is east of the hallway.
How do you go from den to kitchen? A: west, north

This is going to prove difficult for MemNNs because it effectively involves search.
What models could we try?

- Classic NLP cascade e.g. SVM-struct with bunch of features for subtasks: *(Not End-to-End)*
- N-gram models with SVM-type classifier?
- (LSTM) Recurrent Neural Nets?
- Memory Network variants ... ?
- <Insert your new model here>
End-to-end Memory Network (MemN2N)

- New end-to-end (MemN2N) model (Sukhbaatar ‘15):
  - Reads from memory with soft attention
  - Performs multiple lookups (hops) on memory
  - End-to-end training with backpropagation
  - Only need supervision on the final output

- It is based on “Memory Networks” by [Weston, Chopra & Bordes ICLR 2015] but that had:
  - Hard attention
  - requires explicit supervision of attention during training
  - Only feasible for simple tasks
MemN2N architecture

Memory Module

Controller module

Input

\{\vec{m}_1, \vec{m}_2, ..., \vec{m}_N\}

Read

Addressing

\vec{u}_1

\vec{u}_2

\vec{u}_3

Output

supervision

Memory vectors (unordered)

Internal state vector
Memory Module

- Memory vectors
- Attention weights / Soft address
- Dot Product
- Softmax
- Weighted Sum
- \( \sum_{i} p_i \vec{m}_i \)
- To controller (added to controller state)
- Addressing signal (controller state vector)
Question & Answering

Memory Module

Input story:
1: Sam moved to garden
2: Sam went to kitchen
3: Sam drops apple there

Question:
Where is Sam?

Answer:
kitchen

$0.1\vec{m}_1 + 0.7\vec{m}_2 + 0.2\vec{m}_3$

Dot product + softmax

Weighted Sum

Controller

$\vec{u}_1$

$\vec{u}_2$
Memory Vectors

E.g.) constructing memory vectors with Bag-of-Words (BoW)

1. Embed each word

2. Sum embedding vectors

"Sam drops apple" → $\vec{v}_{\text{Sam}} + \vec{v}_{\text{drops}} + \vec{v}_{\text{apple}}$

Embedding Vectors
Positional Encoding of Words

Representation of inputs and memories could use all kinds of encodings: bag of words, RNN style reading at word or character level, etc.

We also built a positional encoding variant: Words are represented by vectors as before. But instead of a bag, position is modeled by a multiplicative term on each word vector with weights depending on the position in the sentence.
<table>
<thead>
<tr>
<th>TASK</th>
<th>N-grams</th>
<th>LSTMs</th>
<th>MemN2N</th>
<th>Memory Networks</th>
<th>StructSVM + coref + srl</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1. Single supporting fact</td>
<td>36</td>
<td>50</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T2. Two supporting facts</td>
<td>2</td>
<td>20</td>
<td>87</td>
<td>PASS</td>
<td>74</td>
</tr>
<tr>
<td>T3. Three supporting facts</td>
<td>7</td>
<td>20</td>
<td>60</td>
<td>PASS</td>
<td>17</td>
</tr>
<tr>
<td>T4. Two arguments relations</td>
<td>50</td>
<td>61</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T5. Three arguments relations</td>
<td>20</td>
<td>70</td>
<td>87</td>
<td>PASS</td>
<td>83</td>
</tr>
<tr>
<td>T6. Yes/no questions</td>
<td>49</td>
<td>48</td>
<td>92</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T7. Counting</td>
<td>52</td>
<td>49</td>
<td>83</td>
<td>85</td>
<td>69</td>
</tr>
<tr>
<td>T8. Sets</td>
<td>40</td>
<td>45</td>
<td>90</td>
<td>91</td>
<td>70</td>
</tr>
<tr>
<td>T9. Simple negation</td>
<td>62</td>
<td>64</td>
<td>87</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T10. Indefinite knowledge</td>
<td>45</td>
<td>44</td>
<td>85</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T11. Basic coreference</td>
<td>29</td>
<td>72</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T12. Conjunction</td>
<td>9</td>
<td>74</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T13. Compound coreference</td>
<td>26</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T14. Time reasoning</td>
<td>19</td>
<td>27</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T15. Basic deduction</td>
<td>20</td>
<td>21</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>T16. Basic induction</td>
<td>43</td>
<td>23</td>
<td>PASS</td>
<td>PASS</td>
<td>24</td>
</tr>
<tr>
<td>T17. Positional reasoning</td>
<td>46</td>
<td>51</td>
<td>49</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>T18. Size reasoning</td>
<td>52</td>
<td>52</td>
<td>89</td>
<td>PASS</td>
<td>62</td>
</tr>
<tr>
<td>T19. Path finding</td>
<td>0</td>
<td>8</td>
<td>7</td>
<td>36</td>
<td>49</td>
</tr>
<tr>
<td>T20. Agent’s motivation</td>
<td>76</td>
<td>91</td>
<td>PASS</td>
<td>PASS</td>
<td>PASS</td>
</tr>
</tbody>
</table>
Attention during memory lookups

Samples from toy QA tasks

<table>
<thead>
<tr>
<th>Story (1: 1 supporting fact)</th>
<th>Support</th>
<th>Hop 1</th>
<th>Hop 2</th>
<th>Hop 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel went to the bathroom.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Mary travelled to the hallway.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>John went to the bedroom.</td>
<td>0.37</td>
<td>0.02</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>John travelled to the bathroom.</td>
<td>yes</td>
<td>0.60</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Mary went to the office.</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*Where is John? Answer: bathroom*  
*Prediction: bathroom*

<table>
<thead>
<tr>
<th>Story (2: 2 supporting facts)</th>
<th>Support</th>
<th>Hop 1</th>
<th>Hop 2</th>
<th>Hop 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>John dropped the milk.</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>John took the milk there.</td>
<td>yes</td>
<td>0.88</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sandra went back to the bathroom.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>John moved to the hallway.</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Mary went to the bedroom.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*Where is the milk? Answer: hallway*  
*Prediction: hallway*

<table>
<thead>
<tr>
<th>Story (16: basic induction)</th>
<th>Support</th>
<th>Hop 1</th>
<th>Hop 2</th>
<th>Hop 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian is a frog.</td>
<td>yes</td>
<td>0.00</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>Lily is gray.</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Brian is yellow.</td>
<td>yes</td>
<td>0.07</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Julius is green.</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Greg is a frog.</td>
<td>yes</td>
<td>0.76</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*What color is Greg? Answer: yellow*  
*Prediction: yellow*

<table>
<thead>
<tr>
<th>Story (18: size reasoning)</th>
<th>Support</th>
<th>Hop 1</th>
<th>Hop 2</th>
<th>Hop 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>The suitcase is bigger than the chest.</td>
<td>yes</td>
<td>0.00</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>The box is bigger than the chocolate.</td>
<td>0.04</td>
<td>0.05</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>The chest is bigger than the chocolate.</td>
<td>yes</td>
<td>0.17</td>
<td>0.07</td>
<td>0.90</td>
</tr>
<tr>
<td>The chest fits inside the container.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>The chest fits inside the box.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*Does the suitcase fit in the chocolate? Answer: no*  
*Prediction: no*

---

Table 2: The perplexity on the test sets of Penn Treebank and Text8 corpora. Note that increasing the number of memory hops improves performance.

Table 3: The accuracy on the test sets of bAbI tasks.  
20 bAbI Tasks

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Acc</th>
<th>Failed tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>MemNN</td>
<td>93.3%</td>
<td>4</td>
</tr>
<tr>
<td>LSTM</td>
<td>49%</td>
<td>20</td>
</tr>
<tr>
<td>MemN2N</td>
<td>74.82%</td>
<td>17</td>
</tr>
<tr>
<td>MemN2N</td>
<td>84.4%</td>
<td>11</td>
</tr>
<tr>
<td>MemN2N</td>
<td>87.6%</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 2: Example predictions on the QA tasks of [21]. 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.

Figure 3: Average activation weight of memory positions during 6 memory hops. White color indicates where the model is attending during the $k$th hop. For clarity, each row is normalized to have maximum value of 1. A model is trained on (left) Penn Treebank and (right) Text8 dataset.

---

Language Modeling Experiments

The goal in language modeling is to predict the next word in a text sequence given the previous words $x$. We now explain how our model can easily be applied to this task.

We now operate on word level, as opposed to the sentence level. Thus the previous $N$ words in the sequence (including the current) are embedded into memory separately. Each memory cell holds only a single word, so there is no need for the BoW or linear mapping representations used in the QA tasks. We employ the temporal embedding approach of Section 4.1.

Since there is no longer any question, $q$ in Fig. 1 is fixed to a constant vector 0.1 (without embedding). The output softmax predicts which word in the vocabulary (of size $V$) is next in the sequence. A cross-entropy loss is used to train model by backpropagating the error through multiple...
So we still fail on some tasks....

.. and we could also make more tasks that we fail on!

Our hope is that a feedback loop of:

1. Developing tasks that break models, and
2. Developing models that can solve tasks

... leads in a fruitful research direction....
How about on real data?

- Toy AI tasks are important for developing innovative methods.
- But they do not give all the answers.

- How do these models work on real data?
  - Classic Language Modeling (Penn TreeBank, Text8)
  - Story understanding (Children’s Book Test, News articles)
  - Open Question Answering (WebQuestions, WikiQA)
  - Goal-Oriented Dialog and Chit-Chat (Movie Dialog, Ubuntu)
Language Modeling

The goal is to predict the next word in a text sequence given the previous words. Results on the Penn Treebank and Text8 (Wikipedia-based) corpora.

<table>
<thead>
<tr>
<th></th>
<th>Penn Tree</th>
<th>Text8</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>129</td>
<td>184</td>
</tr>
<tr>
<td>LSTM</td>
<td>115</td>
<td>154</td>
</tr>
<tr>
<td>MemN2N</td>
<td>121</td>
<td>187</td>
</tr>
<tr>
<td>2 hops</td>
<td>118</td>
<td>154</td>
</tr>
<tr>
<td>7 hops</td>
<td>111</td>
<td>147</td>
</tr>
</tbody>
</table>

Test perplexity

Hops vs. Attention: Average over (PTB)

Average over (Text8)
Language Modeling

The goal is to predict the next word in a text sequence given the previous words. Results on the Penn Treebank and Text8 (Wikipedia-based) corpora.

<table>
<thead>
<tr>
<th></th>
<th>Penn Tree</th>
<th>Text8</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>129</td>
<td>184</td>
</tr>
<tr>
<td>LSTM</td>
<td>115</td>
<td>154</td>
</tr>
<tr>
<td>MemN2N 2 hops</td>
<td>121</td>
<td>187</td>
</tr>
<tr>
<td>5 hops</td>
<td>118</td>
<td>154</td>
</tr>
<tr>
<td>7 hops</td>
<td>111</td>
<td>147</td>
</tr>
</tbody>
</table>

Test perplexity

MemNNs are in the same ballpark as LSTMs.

Hypothesis: many words (e.g. syntax words) don’t actually need really long term context, and so memNNs don’t help there.

Maybe MemNNs could eventually help more on things like nouns/entities?
Children books understanding

New dataset based on 118 children books from project Gutenberg
MemNNs for story understanding

<table>
<thead>
<tr>
<th>m₀</th>
<th>NULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Why, what are YOUR shoes done with?&quot; said the Gryphon</td>
<td></td>
</tr>
<tr>
<td>&quot;I mean, what makes them so shiny?&quot;</td>
<td></td>
</tr>
<tr>
<td>Alice looked down at them, and considered a little before she gave her answer.</td>
<td></td>
</tr>
</tbody>
</table>

| mₙ   | ..... |

`Why, what are YOUR shoes done with?‘ said the Gryphon. ‘I mean, what makes them so shiny?‘ Alice looked down at them, and considered a little before she gave her answer. ‘They’re done with blacking, I believe.‘ ‘Boots and shoes under the sea,‘ the Gryphon went on in a deep voice, ‘are done with a whiting. Now you know.‘ ‘And what are they made of?‘ Alice asked in a tone of great curiosity..’

Cands: Gryphon | Alice | King | Queen | ...

`Soles and eels, of course,‘ the _____ replied rather impatiently: ‘any shrimp could have told you that
MemNNs for story understanding

<table>
<thead>
<tr>
<th>m₀</th>
<th>NULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why</td>
<td>What</td>
</tr>
<tr>
<td>are</td>
<td>your</td>
</tr>
</tbody>
</table>

Size of memories:
1) Sentence?
2) Window?
3) Word?

Memory reads and stores story

`' Why, what are YOUR shoes done with?' said the Gryphon. 'I mean, what makes them so shiny?' Alice looked down at them, and considered a little before she gave her answer. 'They're done with blacking, I believe.' 'Boots and shoes under the sea,' the Gryphon went on in a deep voice, 'are done with a whiting. Now you know.' 'And what are they made of?' 'Alice asked in a tone of great curiosity ..'`

Cands: Gryphon | Alice | King | Queen | ...

`' Soles and eels, of course,' the _____ replied rather impatiently: 'any shrimp could have told you that`
Self-Supervision Memory Network

Two tricks together that make things work a bit better:

1) Bypass module

Instead of the last output module being a linear layer from the output of the memory, assume the answer is one of the memories. Sum the scores of identical memories.

2) Self-Supervision

We know what the right answer is on the training data, so just directly train that memories containing the answer word to be supporting facts (have high probability).
### Results on Children’s Book Test

<table>
<thead>
<tr>
<th>Methods</th>
<th>Named Entities</th>
<th>Common Nouns</th>
<th>Verbs</th>
<th>Prepositions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans (query) (*)</td>
<td>0.520</td>
<td>0.644</td>
<td>0.716</td>
<td>0.676</td>
</tr>
<tr>
<td>Humans (context+query) (*)</td>
<td><strong>0.816</strong></td>
<td><strong>0.816</strong></td>
<td><strong>0.828</strong></td>
<td><strong>0.708</strong></td>
</tr>
<tr>
<td>Maximum frequency (corpus)</td>
<td>0.120</td>
<td>0.158</td>
<td>0.373</td>
<td>0.315</td>
</tr>
<tr>
<td>Maximum frequency (context)</td>
<td>0.335</td>
<td>0.281</td>
<td>0.285</td>
<td>0.275</td>
</tr>
<tr>
<td>Sliding window</td>
<td>0.168</td>
<td>0.196</td>
<td>0.182</td>
<td>0.101</td>
</tr>
<tr>
<td>Word distance model</td>
<td>0.398</td>
<td>0.364</td>
<td>0.380</td>
<td>0.237</td>
</tr>
<tr>
<td>Kneser-Ney language model</td>
<td>0.390</td>
<td>0.544</td>
<td>0.778</td>
<td>0.768</td>
</tr>
<tr>
<td>Kneser-Ney language model + cache</td>
<td>0.439</td>
<td>0.577</td>
<td>0.772</td>
<td>0.679</td>
</tr>
<tr>
<td>Embedding Model (context+query)</td>
<td>0.253</td>
<td>0.259</td>
<td>0.421</td>
<td>0.315</td>
</tr>
<tr>
<td>Embedding Model (query)</td>
<td>0.351</td>
<td>0.400</td>
<td>0.614</td>
<td>0.535</td>
</tr>
<tr>
<td>Embedding Model (window)</td>
<td>0.362</td>
<td>0.415</td>
<td>0.637</td>
<td>0.589</td>
</tr>
<tr>
<td>Embedding Model (window+position)</td>
<td>0.402</td>
<td>0.506</td>
<td>0.736</td>
<td>0.670</td>
</tr>
<tr>
<td>LSTMs (query)</td>
<td>0.408</td>
<td>0.541</td>
<td>0.813</td>
<td>0.802</td>
</tr>
<tr>
<td>LSTMs (context+query)</td>
<td>0.418</td>
<td>0.560</td>
<td><strong>0.818</strong></td>
<td><strong>0.791</strong></td>
</tr>
<tr>
<td>Contextual LSTMs (window context)</td>
<td>0.436</td>
<td>0.582</td>
<td>0.805</td>
<td><strong>0.806</strong></td>
</tr>
<tr>
<td>MemNNS (lexical memory)</td>
<td>0.431</td>
<td>0.562</td>
<td>0.798</td>
<td>0.764</td>
</tr>
<tr>
<td>MemNNS (window memory)</td>
<td>0.493</td>
<td>0.554</td>
<td>0.692</td>
<td>0.674</td>
</tr>
<tr>
<td>MemNNS (sentential memory + PE)</td>
<td>0.318</td>
<td>0.305</td>
<td>0.502</td>
<td>0.326</td>
</tr>
<tr>
<td>MemNNS (window memory + self-sup.)</td>
<td><strong>0.666</strong></td>
<td><strong>0.630</strong></td>
<td>0.690</td>
<td>0.703</td>
</tr>
</tbody>
</table>
### Question Answering on New’s Articles

We evaluate our models on the data from:

**“Teaching Machines to Read and Comprehend”**

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom

<table>
<thead>
<tr>
<th>Original Version</th>
<th>Anonymised Version</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context</strong></td>
<td></td>
</tr>
</tbody>
</table>
| The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.”  

  ...                                                                                   | the ent381 producer allegedly struck by ent212 will not press charges against the “ent153” host, his lawyer said friday. ent212, who hosted one of the most-watched television shows in the world, was dropped by the ent381 Wednesday after an internal investigation by the ent180 broadcaster found he had subjected producer ent193 “to an unprovoked physical and verbal attack.”  

  ...                                                                                   |
| **Query**                                                                         |                                                                                  |
| Producer X will not press charges against Jeremy Clarkson, his lawyer says.         | producer X will not press charges against ent212, his lawyer says.               |
| **Answer**                                                                        |                                                                                  |
| Oisin Tymon                                                                       | ent193                                                                           |
## Results on CNN QA dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAXIMUM FREQUENCY (ARTICLE)</strong>(*)</td>
<td>0.305</td>
<td>0.332</td>
</tr>
<tr>
<td><strong>SLIDING WINDOW</strong></td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td><strong>WORD DISTANCE MODEL</strong>(*)</td>
<td>0.505</td>
<td>0.509</td>
</tr>
<tr>
<td><strong>DEEP LSTMs (ARTICLE+QUERY)</strong>(*)</td>
<td>0.550</td>
<td>0.570</td>
</tr>
<tr>
<td><strong>CONTEXTUAL LSTMs (“ATTENTIVE READER”)</strong>(*)</td>
<td>0.616</td>
<td>0.630</td>
</tr>
<tr>
<td><strong>CONTEXTUAL LSTMs (“IMPATIENT READER”)</strong>(*)</td>
<td>0.618</td>
<td>0.638</td>
</tr>
<tr>
<td><strong>MEMNNs (WINDOW MEMORY)</strong></td>
<td>0.580</td>
<td>0.606</td>
</tr>
<tr>
<td><strong>MEMNNs (WINDOW MEMORY + SELF-SUP.)</strong></td>
<td>0.634</td>
<td>0.668</td>
</tr>
<tr>
<td><strong>MEMNNs (WINDOW MEMORY + ENSEMBLE)</strong></td>
<td>0.612</td>
<td>0.638</td>
</tr>
<tr>
<td><strong>MEMNNs (WINDOW MEMORY + SELF-SUP. + ENSEMBLE)</strong></td>
<td>0.649</td>
<td>0.684</td>
</tr>
<tr>
<td><strong>MEMNNs (WINDOW + SELF-SUP. + ENSEMBLE + EXCLUD. COOCURRENCES)</strong></td>
<td><strong>0.662</strong></td>
<td><strong>0.694</strong></td>
</tr>
</tbody>
</table>

Table 3: **Results on CNN QA.** (*)Results taken from Hermann et al. (2015).
Latest *Fresh* Results

- Our best results:  
  - QACNN: 69.4  
  - CBT-NE: 66.6  
  - CBT-V: 63.0

- Text Understanding with the Attention Sum Reader Network. Kadlec et al. (4 Mar ‘16)  
  - QACNN: 75.4  
  - CBT-NE: 71.0  
  - CBT-CN: 68.9  
  *Uses RNN style encoding of words + bypass module + 1 hop*

- Iterative Alternating Neural Attention for Machine Reading. Sordoni et al. (7 Jun ’16)  
  - QACNN: 76.1  
  - CBT-NE: 72.0  
  - CBT-CN: 71.0

- Natural Language Comprehension with the EpiReader. Trischler et al. (7 Jun ’16)  
  - QACNN: 74.0  
  - CBT-NE: 71.8  
  - CBT-CN: 70.6

- Gated-Attention Readers for Text Comprehension. Dhingra et al. (5 Jun ’16)  
  - QACNN: 77.4  
  - CBT-NE: 71.9  
  - CBT-CN: 69.  
  *Uses RNN style encoding of words + bypass module + multiplicative combination of query + multiple hops*
Large Scale QA

Memory reads and stores Freebase

Freebase

22 M facts
5 M entities

Read Module looks for 1 sup. fact among a subset: uses hashing/string matching for fast lookup.

Who created Gollum from The Hobbit?

R Module returns the object

Gollum character created by JRR Tolkien
JRR Tolkien place of birth Bloemfontein
Bloemfontein contained by South Africa
The Hobbit directed by Peter Jackson
Facebook Inc founded by Mark Zuckerberg

JRR Tolkien
WebQuestions & SimpleQuestions

- Decent results on WebQuestions, a popular QA task:

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
<td>1.9</td>
</tr>
<tr>
<td>Bordes et al., 2014b</td>
<td>29.7</td>
</tr>
<tr>
<td>Berant et al., 2013</td>
<td>31.3</td>
</tr>
<tr>
<td>Bordes et al., 2014a</td>
<td>39.2</td>
</tr>
<tr>
<td>Berant and Liang, 2014</td>
<td>39.9</td>
</tr>
<tr>
<td>Yang et al., 2014</td>
<td>41.3</td>
</tr>
<tr>
<td>MemNN</td>
<td>42.2</td>
</tr>
</tbody>
</table>


- However now beaten by many results, especially (Yih et al. ACL ‘15) that achieves **52.5**! Several hand engineered features are used in that case. Note WebQuestions is very small (4k train+valid).
Recent Work: New Models for QA on documents

What year was the movie Blade Runner released?

1982
Recent Work: New Models for QA on documents

WikiQA Results

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Cnt</td>
<td>0.4891</td>
<td>0.4924</td>
</tr>
<tr>
<td>Wgt Word Cnt</td>
<td>0.5099</td>
<td>0.5132</td>
</tr>
<tr>
<td>2-gram CNN (Yang et al., 2015)</td>
<td>0.6520</td>
<td>0.6652</td>
</tr>
<tr>
<td>AP-CNN (Santos et al., 2016)</td>
<td>0.6886</td>
<td>0.6957</td>
</tr>
<tr>
<td>Attentive LSTM (Miao et al., 2015)</td>
<td>0.6886</td>
<td>0.7069</td>
</tr>
<tr>
<td>Attentive CNN (Yin and Schütze, 2015)</td>
<td>0.6921</td>
<td>0.7108</td>
</tr>
<tr>
<td>L.D.C. (Wang et al., 2016)</td>
<td>0.7058</td>
<td>0.7226</td>
</tr>
<tr>
<td>Memory Network</td>
<td>0.5170</td>
<td>0.5236</td>
</tr>
<tr>
<td>Key-Value Memory Network</td>
<td><strong>0.7069</strong></td>
<td><strong>0.7265</strong></td>
</tr>
</tbody>
</table>

What year was the movie Blade Runner released?

Wikipedia Entry: Blade Runner
Blade Runner is a dystopian science fiction film starring Harrison Ford, Rutger Hauer, Sean Young, and Edward James Olmos. The screenplay, written by Hampton Fancher and D. P. Scott, is a modified film adaptation of the 1968 novel “Do Androids Dream of Electric Sheep?” by Philip K. Dick. The film depicts a dystopian Los Angeles in November 1982 in which genetically engineered replicants, which are visually indistinguishable from adult humans, are manufactured by the powerful Tyrell Corporation as...
How about on large scale dialog data? With multiple exchanges?

- Everything we showed so far was question answering potentially with long-term context.

- We have also built a Movie Dialog Dataset
  Closed, but large, domain about movies (75k entities, 3.5M ex).
  - Ask facts about movies?
  - Ask for opinions (recommendations) about movies?
  - Dialog combining facts and opinions?
  - General chit-chat about movies (statements not questions)?

And... combination of all above in one end-to-end model.
Recent Work: Combines QA with Dialog Tasks
Dodge et al. “Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems.” ICLR ‘16

(Dialog 1) QA: facts about movies
Sample input contexts and target replies (in red) from Dialog Task 1:

What movies are about open source? Revolution OS
Ruggero Raimondi appears in which movies? Carmen
Can you name a film directed by Stuart Ortiz? Grave Encounters
Who directed the film White Elephant? Pablo Trapero
What is the genre of the film Dial M for Murder? Thriller, Crime
What language is Whity in? German

(Dialog 2) Recs: movie recommendations
Sample input contexts and target replies (in red) from Dialog Task 2:

Some movies I like are Heat, Kids, Fight Club, Shaun of the Dead, The Avengers, Skyfall, and Jurassic Park. Can you suggest something else I might like? Ocean's Eleven

(Dialog 3) QA+Recs: combination dialog
Sample input contexts and target replies (in red) from Dialog Task 3:

I loved Billy Madison, Blades of Glory, Bio-Dome, Clue, and Happy Gilmore. I’m looking for a Music movie. School of Rock
What else is that about? Music, Musical, Jack Black, school, teacher, Richard Linklater, rock, guitar
I like rock and roll movies more. Do you know anything else? Little Richard

(Dialog 4) Reddit: real dialog
Sample input contexts and target replies (in red) from Dialog Task 4:

I think the Terminator movies really suck, I mean the first one was kinda ok, but after that they got really cheesy. Even the second one which people somehow think is great. And after that... forgeddabotit.
C’mon the second one was still pretty cool.. Army was still so badass, as was Sararah Connor’s character.. and the way they blended real action and effects was perhaps the last of its kind...
(Dialog 1) QA: facts about movies

Sample input contexts and target replies (in red) from Dialog Task 1:

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What movies are about open source?</td>
<td>Revolution OS</td>
</tr>
<tr>
<td>Ruggero Raimondi appears in which movies?</td>
<td>Carmen</td>
</tr>
<tr>
<td>Can you name a film directed by Stuart Ortiz?</td>
<td>Grave Encounters</td>
</tr>
<tr>
<td>Who directed the film White Elephant?</td>
<td>Pablo Trapero</td>
</tr>
<tr>
<td>What is the genre of the film Dial M for Murder?</td>
<td>Thriller, Crime</td>
</tr>
<tr>
<td>What language is Whity in?</td>
<td>German</td>
</tr>
</tbody>
</table>

Some movies I like are Heat, Kids, Fight Club, Shaun of the Dead, The Avengers, Skyfall, and Jurassic Park. Can you suggest something else I might like? **Ocean's Eleven**
I loved Billy Madison, Blades of Glory, Bio-Dome, Clue, and Happy Gilmore. I'm looking for a Music movie. School of Rock
What else is that about? Music, Musical, Jack Black, school, teacher, Richard Linklater, rock, guitar
I like rock and roll movies more. Do you know anything else? Little Richard
I think the Terminator movies really suck, I mean the first one was kinda ok, but after that they got really cheesy. Even the second one which people somehow think is great. And after that... forgeddabotit.
C’mon the second one was still pretty cool.. Arny was still so badass, as was Sararah Connor’s character.. and the way they blended real action and effects was perhaps the last of its kind...
Memory Network:

Memories \( h_i \)

- **Shaolin Soccer** written by Stephen Chow
- **Shaolin Soccer** starred actors Stephen Chow
- **Shaolin Soccer** release year 2001
- **Shaolin Soccer** has genre comedy
- **Shaolin Soccer** has tags martial arts, kung fu soccer, stephen chow
- **Kung Fu Hustle** directed by Stephen Chow
- **Kung Fu Hustle** written by Stephen Chow
- **Kung Fu Hustle** starred actors Stephen Chow
- **Kung Fu Hustle** has genre comedy action
- **Kung Fu Hustle** has imdb votes famous
- **Kung Fu Hustle** has tags comedy, action, martial arts, kung fu, china, soccer, hong kong, stephen chow
- The God of Cookery directed by Stephen Chow
- The God of Cookery written by Stephen Chow
- The God of Cookery starred actors Stephen Chow
- The God of Cookery has tags hong kong stephen chow
- From Beijing with Love directed by Stephen Chow
- From Beijing with Love written by Stephen Chow
- From Beijing with Love starred actors Stephen Chow, Anita Yuen

...<and more>...

Short-Term \( c^u_1 \)

1) I’m looking a fun comedy to watch tonight, any ideas?
2) Have you seen **Shaolin Soccer**? That was zany and great.. really funny but in a whacky way.

Input \( c^u_2 \)

3) Yes! **Shaolin Soccer** and **Kung Fu Hustle** are so good I really need to find some more Stephen Chow films I feel like there is more awesomeness out there that I haven’t discovered yet...
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>QA Task (hits@1)</th>
<th>Recs Task (hits@100)</th>
<th>QA+Recs Task (hits@10)</th>
<th>Reddit Task (hits@10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA System (Bordes et al., 2014)</td>
<td>90.7</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SVD</td>
<td>N/A</td>
<td>19.2</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>IR</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>23.7</td>
</tr>
<tr>
<td>LSTM</td>
<td>6.5</td>
<td>27.1</td>
<td>19.9</td>
<td>11.8</td>
</tr>
<tr>
<td>Supervised Embeddings</td>
<td>50.9</td>
<td>29.2</td>
<td>65.9</td>
<td>27.6</td>
</tr>
<tr>
<td>MemN2N</td>
<td>79.3</td>
<td>28.6</td>
<td>81.7</td>
<td>29.2</td>
</tr>
<tr>
<td>Joint Supervised Embeddings</td>
<td>43.6</td>
<td>28.1</td>
<td>58.9</td>
<td>14.5</td>
</tr>
<tr>
<td>Joint MemN2N</td>
<td>83.5</td>
<td>26.5</td>
<td>78.9</td>
<td>26.6</td>
</tr>
</tbody>
</table>
Ubuntu Data

Dialog dataset: Ubuntu IRC channel logs, users ask questions about issues they are having with Ubuntu and get answers by other users. (Lowe et al., ’15)

<table>
<thead>
<tr>
<th>METHODS</th>
<th>VALIDATION (HITS@1)</th>
<th>TEST (HITS@1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>N/A</td>
<td>48.81</td>
</tr>
<tr>
<td>RNN</td>
<td>N/A</td>
<td>37.91</td>
</tr>
<tr>
<td>LSTM</td>
<td>N/A</td>
<td>55.22</td>
</tr>
<tr>
<td>MemN2N 1-HOP</td>
<td></td>
<td>57.23</td>
</tr>
<tr>
<td>MemN2N 2-HOPS</td>
<td></td>
<td>64.28</td>
</tr>
<tr>
<td>MemN2N 3-HOPS</td>
<td></td>
<td>64.31</td>
</tr>
<tr>
<td>MemN2N 4-HOPS</td>
<td></td>
<td>64.01</td>
</tr>
</tbody>
</table>

Table 7: Ubuntu Dialog Corpus results. The evaluation is retrieval-based, similar to that of Reddit (Task 4). For each dialog, the correct answer is mixed among 10 random candidates; Hits@1 (in %) are reported. Methods with † have been ran by Lowe et al. (2015).

Best results currently reported:
Sentence Pair Scoring: Towards Unified Framework for Text Comprehension
Petr Baudiš, Jan Pichl, Tomáš Vyskočil, Jan Šedivý
RNN-CNN combo model: 67.2
Next Steps

Artificial tasks to help design new methods:

- New methods that succeed on all bAbI tasks?
- Make more bAbI tasks to check other skills.

Real tasks to make sure those methods are actually useful:

- Sophisticated reasoning on bAbI tasks doesn’t always happen as clearly on real data.. Why? Fix!
- Models that work jointly on all tasks so far built.

Dream: can learn from very weak supervision:

We would like to learn in an environment just by communicating with other agents / humans, as well as seeing other agents communicating + acting in the environment.

E.g. a baby talking to its parents, and seeing them talk to each other.
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.

Where is Mary?  A: playground
No, that's incorrect.

Where is John?  A: bathroom
Yes, that's right!

If you can predict this, you are most of the way to knowing how to answer correctly.
Human Responses Give Lots of Info

Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.

Where is Mary?  A: playground
No, the answer is kitchen.

Where is John?  A: bathroom
Yes, that's right!

Much more signal than just “No” or zero reward.
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A: playground
No, she’s in the kitchen.

If you can predict this, you are most of the way to knowing how to answer correctly.

FAIR: paper / data / code

Papers:
- bAbI tasks: arxiv.org/abs/1502.05698
- Memory Networks: http://arxiv.org/abs/1410.3916
- End-to-end Memory Networks: http://arxiv.org/abs/1503.08895
- Large-scale QA with MemNNs: http://arxiv.org/abs/1506.02075

Data:
- bAbI tasks: fb.ai/babi
- SimpleQuestions dataset (100k questions): fb.ai/babi
- Children’s Book Test dataset: fb.ai/babi
- Movie Dialog Dataest: fb.ai/babi

Code:
- Memory Networks: https://github.com/facebook/MemNN
- Simulation tasks generator: https://github.com/facebook/bAbI-tasks
RAM Issues

- How to decide what to write and what not to write in the memory?
- How to represent knowledge to be stored in memories?
- Types of memory (arrays, stacks, or stored within weights of model), when they should be used, and how can they be learnt?
- How to do fast retrieval of relevant knowledge from memories when the scale is huge?
- How to build hierarchical memories, e.g. multiscale attention?
- How to build hierarchical reasoning, e.g. composition of functions?
- How to incorporate forgetting/compression of information?
- How to evaluate reasoning models? Are artificial tasks a good way? Where do they break down and real tasks are needed?
- Can we draw inspiration from how animal or human memories work?
Thanks!