Reasoning with Memory Augmented Neural Networks for Language Comprehension, ICLR17

Tsenduren Munkhdalai & Hong Yu

UMass Medical School

ICLR 2017
Presenter: Jack Lanchantin
Outline

1. Intro

2. Hypothesis Testing Model for Cloze-Type QA
   - Memory Initialization
   - Hypothesis-Test Loop
   - Ending The Hypothesis-Test Loop

3. Experiments

4. Conclusions
Hypothesis Testing

- In QA, a hypothesis is formed by regressing the original query, and then the hypothesis is tested against reality (support set).
- If the model is satisfied with the current hypothesis, the reasoning process is stopped and the answer is found. Else, reformulate hypothesis until the answer is found.
Context
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.” …

Query
Producer X will not press charges against Jeremy Clarkson, his lawyer says.

Answer
Oisin Tymon
Many similar models for this task: Attentive Reader (Hermann et. al. 2015, Memory Networks (Sukhbaatar et. al. 2015), EpiReader (Trischler et al. 2016)

Previous models designed with the predefined number of computational hops for inference, but different types of questions require different types of reasoning
Outline

1 Intro

2 Hypothesis Testing Model for Cloze-Type QA
   - Memory Initialization
   - Hypothesis-Test Loop
   - Ending The Hypothesis-Test Loop

3 Experiments

4 Conclusions
Memory Initialization

Query Memory:  \( M_0^q = BiLSTM^q(Q) \in \mathbb{R}^{k \times |Q|} \) \hspace{1cm} (1)

Document Memory:  \( M^d = BiLSTM^d(D) \in \mathbb{R}^{k \times |D|} \) \hspace{1cm} (2)
Outline

1 Intro

2 Hypothesis Testing Model for Cloze-Type QA
   - Memory Initialization
   - Hypothesis-Test Loop
   - Ending The Hypothesis-Test Loop

3 Experiments

4 Conclusions
Updating Memory through Hypothesis-Test Loop

- **Hypothesis Formulation**: In each step, the query memory $M^q$ is updated with content from the document memory to form a new query.

- **Hypothesis Testing**: The new query is then checked against the document facts to make an answer prediction.

- **3 Steps**:
  1. Read
  2. Compose
  3. Write
Given Query state $s^q_t \in \mathbb{R}^k$, and Document state $s^d_t \in \mathbb{R}^k$:

Read vector ($k$): $r_t = LSTM^r([s^q_{t-1}; s^d_{t-1}])$ \hspace{1cm} (3)

Q Alignment ($|Q|$): $l^q_t = r^T_t M^q_t$ \hspace{1cm} (4)

New Q State ($k$): $s^q_t = \text{softmax}(l^q_t)^T M^q_{t-1}$ \hspace{1cm} (5)

Q Memory Key ($|Q|$): $z^q_t = \text{sigmoid}(l^q_t)$ \hspace{1cm} (6)

D Alignment ($|D|$): $l^d_t = s^q_t T M^d$ \hspace{1cm} (7)

New D State ($k$): $s^d_t = \text{softmax}(l^d_t)^T M^d$ \hspace{1cm} (8)
2. Compose

Combines the current query state $s^q_t$, document state $s^d_t$, and hidden state of read module $r_t$, resulting in single vector $c_t$

\[ c_t \in \mathbb{R}^k = MLP^c(s^q_t, s^d_t, r_t) \quad (9) \]
3. Write

Takes outputs $z^q_t$ (Q Memory Key), and $s^d_t$ (Document State) from the read module and updates the query memory:

$$M^q_t = M^q_{t-1}z^q_t + s^d_t(1 - z^q_t)$$ (10)
Outline

1. Intro

2. Hypothesis Testing Model for Cloze-Type QA
   - Memory Initialization
   - Hypothesis-Test Loop
   - Ending The Hypothesis-Test Loop

3. Experiments

4. Conclusions
Ending The Hypothesis-Test Loop

- Need to decide when to stop the read-compose-write loop and answer
- 2 strategies are explored:
  1. Query-Gating
  2. Adaptive Computation
Query Gating

Instead of making a hard decision on halting the loop (i.e. stop reading), the write module performs a word-level query gating as:

- **Write vector (k):** $$w_t = LSTM^w(c_t)$$ (11)
- **Write gate (k):** $$g^q_t = \text{sigmoid}(w_t^TM^q_{t-1})$$ (12)
- **New Query Memory (k × |Q|):** $$M^q_t = M^q_t(1 - g^q_t) + M^q_{t-1}g^q_t$$ (13)
Adaptive Computation

Write module is equipped with a termination, which decides its willingness to continue or finish in each step:

Write vector \( (k) \): \( w_t = LSTM^w(c_t) \) \hspace{1cm} (14)

Termination Score (scalar): \( e_t = \text{sigmoid}(o^T w_t) \) \hspace{1cm} (15)

End probability: \( p_t = e_t \prod_{i=1}^{t-1} (1 - e_i) \), \( p_T = 1 - \sum_{i=1}^{T-1} p_i \) \hspace{1cm} (16)
Answer Prediction

\[ P_t(a|Q, D) = v^T \text{softmax}(l_t^d) \]  \hspace{1cm} (17)

where \( v \in \mathbb{R}^{|D|} \) is a mask denoting the positions of the answer token \( a \) in the document (ones for the token and zeros otherwise). \( l_t^d \) is the query-document alignment score.
Children’s Book Test (CBT)

S: 1 Mr. Cropper was opposed to our hiring you.
2 Not, of course, that he had any personal objection to you, but he is set
against female teachers, and when a Cropper is set there is nothing on earth can
change him.
3 He says female teachers can’t keep order.
4 He ‘s started in with a spite at you on general principles, and the boys know
it.
5 They know he ’ll back them up in secret, no matter what they do, just to prove
his opinions.
6 Cropper is sly and slippery, and it is hard to corner him. ’
7 " Are the boys big?"'
8 queried Esther anxiously.
9 " Yes.
10 Thirteen and fourteen and big for their age.
11 You can’t whip ‘em -- that is the trouble.
12 A man might, but they’d twist you around their fingers.
13 You’ll have your hands full, I’m afraid.
14 But maybe they’ll behave all right after all.
15 Mr. Baxter privately had no hope that they would, but Esther hoped for the
best.
16 She could not believe that Mr. Cropper would carry his prejudices into a
personal application.
17 This conviction was strengthened when he overtook her walking from school the
next day and drove her home.
18 He was a big, handsome man with a very suave, polite manner.
19 He asked interestedly about her school and her work, hoped she was getting on
well, and said he had two young rascals of his own to send soon.
20 Esther felt relieved.

Q: She thought that Mr. ______ had exaggerated matters a little.

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

- 20 consecutive sentences from a childrens story
- 21st sentence in which a word has been deleted
- list of ten choices for the deleted word
- tests story completion rather than reading comprehension
### Table 2: Model comparison on the CBT dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>CBT-NE dev</th>
<th>CBT-NE test</th>
<th>CBT-CN dev</th>
<th>CBT-CN test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (context + query) (Hill et al., 2015)</td>
<td>-</td>
<td>81.6</td>
<td>-</td>
<td>81.6</td>
</tr>
<tr>
<td>LSTMs (context + query) (Hill et al., 2015)</td>
<td>51.2</td>
<td>41.8</td>
<td>62.6</td>
<td>56.0</td>
</tr>
<tr>
<td>MemNNs (window mem. + self-sup.) (Hill et al., 2015)</td>
<td>70.4</td>
<td>66.6</td>
<td>64.2</td>
<td>63.0</td>
</tr>
<tr>
<td>AS Reader (Kadlec et al., 2016)</td>
<td>73.8</td>
<td>68.6</td>
<td>68.8</td>
<td>63.4</td>
</tr>
<tr>
<td>GA Reader (Dhingra et al., 2016)</td>
<td>74.9</td>
<td>69.0</td>
<td>69.0</td>
<td>63.9</td>
</tr>
<tr>
<td>EpiReader (Trischler et al., 2016)</td>
<td>75.3</td>
<td>69.7</td>
<td>71.5</td>
<td>67.4</td>
</tr>
<tr>
<td>IAA Reader (Sordoni et al., 2016)</td>
<td>75.2</td>
<td>68.6</td>
<td>72.1</td>
<td>69.2</td>
</tr>
<tr>
<td>AoA Reader (Cui et al., 2016)</td>
<td>77.8</td>
<td>72.0</td>
<td>72.2</td>
<td>69.4</td>
</tr>
<tr>
<td>MemNN (window mem. + self-sup. + ensemble) (Hill et al., 2015)</td>
<td>70.4</td>
<td>66.6</td>
<td>64.2</td>
<td>63.0</td>
</tr>
<tr>
<td>AS Reader (ensemble) (Kadlec et al., 2016)</td>
<td>74.5</td>
<td>70.6</td>
<td>71.1</td>
<td>68.9</td>
</tr>
<tr>
<td>EpiReader (ensemble) (Trischler et al., 2016)</td>
<td>76.6</td>
<td>71.8</td>
<td>73.6</td>
<td>70.6</td>
</tr>
<tr>
<td>IAA Reader (ensemble) (Sordoni et al., 2016)</td>
<td>76.9</td>
<td>72.0</td>
<td>74.1</td>
<td>71.0</td>
</tr>
<tr>
<td>NSE ($T = 1$)</td>
<td>76.2</td>
<td>71.1</td>
<td>72.8</td>
<td>69.7</td>
</tr>
<tr>
<td>NSE Query Gating ($T = 2$)</td>
<td>76.6</td>
<td>71.5</td>
<td>72.3</td>
<td>70.7</td>
</tr>
<tr>
<td>NSE Query Gating ($T = 6$)</td>
<td>77.0</td>
<td>71.4</td>
<td>73.0</td>
<td><strong>72.0</strong></td>
</tr>
<tr>
<td>NSE Query Gating ($T = 9$)</td>
<td>78.0</td>
<td>72.6</td>
<td>73.5</td>
<td>71.2</td>
</tr>
<tr>
<td>NSE Query Gating ($T = 12$)</td>
<td>77.7</td>
<td>72.2</td>
<td><strong>74.3</strong></td>
<td>71.9</td>
</tr>
<tr>
<td>NSE Adaptive Computation ($T = 2$)</td>
<td>77.1</td>
<td>72.1</td>
<td>72.8</td>
<td>71.2</td>
</tr>
<tr>
<td>NSE Adaptive Computation ($T = 12$)</td>
<td><strong>78.2</strong></td>
<td><strong>73.2</strong></td>
<td>74.2</td>
<td>71.4</td>
</tr>
</tbody>
</table>

CBT-NE: Named Entities, CBT-CN: Common Nouns
Passage: Tottenham won 2-0 at Hapoel Tel Aviv in UEFA Cup action on Thursday night in a defensive struggle which impressed Spurs skipper Robbie Keane. ... Keane scored the first goal at the Bloomfield Stadium with Dimitar Berbatov, who insisted earlier on Thursday he was happy at the London club, heading a second. The 26-year-old Berbatov, who has been the subject of reports linking him with a move had affected his performances ... Spurs manager Juande Ramos has won the last two seasons ...

Question: Tottenham manager Juande Ramos has hinted he will allow XXX to leave if the Bulgaria striker remains unhappy.

Choices: (1) Robbie Keane (2) Dimitar Berbatov
## Table 3: Model comparison on the WDW dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Strict</th>
<th></th>
<th>Relaxed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dev</td>
<td>test</td>
<td>dev</td>
<td>test</td>
</tr>
<tr>
<td>Human (Onishi et al., 2016)</td>
<td>-</td>
<td>84.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Attentive Reader (Hermann et al., 2015)</td>
<td>-</td>
<td>53.0</td>
<td>-</td>
<td>55.0</td>
</tr>
<tr>
<td>AS Reader (Kadlec et al., 2016)</td>
<td>-</td>
<td>57.0</td>
<td>-</td>
<td>59.0</td>
</tr>
<tr>
<td>GA Reader (Dhingra et al., 2016)</td>
<td>-</td>
<td>57.0</td>
<td>-</td>
<td>60.0</td>
</tr>
<tr>
<td>Stanford Attentive Reader (Chen et al., 2016)</td>
<td>-</td>
<td>64.0</td>
<td>-</td>
<td>65.0</td>
</tr>
<tr>
<td>NSE (T = 1)</td>
<td>65.1</td>
<td>65.5</td>
<td>66.4</td>
<td>65.3</td>
</tr>
<tr>
<td>NSE Query Gating (T = 2)</td>
<td>65.4</td>
<td>65.1</td>
<td>65.7</td>
<td>65.5</td>
</tr>
<tr>
<td>NSE Query Gating (T = 6)</td>
<td>65.5</td>
<td>65.7</td>
<td>65.6</td>
<td>65.8</td>
</tr>
<tr>
<td>NSE Query Gating (T = 9)</td>
<td>65.8</td>
<td>65.8</td>
<td>65.8</td>
<td>65.9</td>
</tr>
<tr>
<td>NSE Query Gating (T = 12)</td>
<td>65.2</td>
<td>65.5</td>
<td>65.7</td>
<td>65.4</td>
</tr>
<tr>
<td>NSE Adaptive Computation (T = 2)</td>
<td>65.3</td>
<td>65.4</td>
<td>66.2</td>
<td>66.0</td>
</tr>
<tr>
<td>NSE Adaptive Computation (T = 12)</td>
<td><strong>66.5</strong></td>
<td><strong>66.2</strong></td>
<td><strong>67.0</strong></td>
<td><strong>66.7</strong></td>
</tr>
</tbody>
</table>
Conclusions

- Hypothesis Testing QA
- Query memory gating which prevents from forgetting old query and adaptive computation with termination head