A Neural Network for Factoid Question Answering over Paragraphs

code & data: http://cs.umd.edu/~miyery/qblearn

Mohit Iyyer1, Jordan Boyd-Graber2, Leonardo Claudino1, Richard Socher3, and Hal Daumé III1

University of Maryland, College Park1 University of Colorado, Boulder2 Stanford University3

THE TASK: QUIZ BOWL

Quiz bowl is a trivia game where players are read paragraph-length questions and can “buzz in” at any point during the question.

Q: He left unfinished a novel whose title character forges his father’s signature to get out of school and avoids the draft by feigning desire to join . A more famous work by this author tells of the rise and fall of the character forges his father’s signature to get .

A: Thomas Mann

WHY IS THIS CHALLENGING?

• Question pyramidality: earlier sentences contain harder clues than later ones
• Early sentences usually contain very few if any named entities indicative of the answer
• Have to decide when to answer the question as well as what answer to give

WHY NOT TRADITIONAL QA?

• IR systems work by querying some large knowledge base for terms similar to those in the query. But what if the query lacks informative terms?
• In such cases, we have to model the compositionality of the query.

CONTRIBUTIONS OF OUR WORK

• A dependency-tree recursive neural network model, QANTA, that computes distributed question representations to predict answers.
• QANTA outperforms multiple strong baselines and defeats human quiz bowl players when combined with IR methods.

EXPERIMENTAL MODELS:

• BoW, BoW-DT – unigram bag-of-words logistic regression baseline
• IR-QB, IR-WIKI – uses Whoosh, an IR engine, to search a knowledge base of training QA pairs and Wikipedia with BM-25 term weighting, query expansion, and fuzzy queries.
• QANTA, FIXED QANTA – our DT-RNN model, trained only on QA pairs, vary answer training.
• QANTA + IR-WIKI – combines DT-RNN features + IR scores, our best model

DATA:

• Data was provided by NAQT (naqt.com).
• History dataset: 4,460 questions (16,985 sentences), literature dataset: 5,695 questions (21,549 sentences).

RESULTS:

<table>
<thead>
<tr>
<th>Model</th>
<th>History</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pos 1</td>
<td>Pos 2</td>
</tr>
<tr>
<td>BoW</td>
<td>27.5</td>
<td>51.3</td>
</tr>
<tr>
<td>BoW-DT</td>
<td>35.4</td>
<td>57.7</td>
</tr>
<tr>
<td>IR-QB</td>
<td>37.5</td>
<td>65.9</td>
</tr>
<tr>
<td>FIXED-QANTA</td>
<td>38.3</td>
<td>64.4</td>
</tr>
<tr>
<td>QANTA</td>
<td>47.1</td>
<td>72.1</td>
</tr>
<tr>
<td>QANTA + IR-WIKI</td>
<td>53.7</td>
<td>76.6</td>
</tr>
</tbody>
</table>

HUMAN EVALUATION:

We compare our model to 22 skilled quiz bowl players on both datasets; we beat the average human at history questions.

LEARNING ATTRIBUTES

FUTURE WORK:

• demo QANTA at the 2015 NAQT High School National Championships