Question Difficulty Prediction for READING Problems in Standard Tests

Zhenya Huang†, Qi Liu†*, Enhong Chen†, Hongke Zhao†, Mingyong Gao‡, Si Wei‡, Yu Su‡, Guoping Hu‡
†School of Computer Science and Technology, University of Science and Technology of China
{huangzhy, zhhk}@mail.ustc.edu.cn, {qiliuql, cheneh}@ustc.edu.cn
‡iFLYTEK Research, {mygao2, siwei, gphu}@iflytek.com
§School of Computer Science and Technology, Anhui University, yusu@iflytek.com

Reporter: Zhenya Huang
Date: Feb. 7th, 2017
In widely used standard tests, such as TOEFL, examinees are often allowed to retake tests and choose higher scores for college admission.

- **Fairness requirement**: select test papers with consistent difficulties.
- Test Measurements have attracted much attention.
- **Crucial demand**: question difficulty prediction (QDP)
What is question difficulty?

- Following Educational Psychology, question difficulty refers to the percentage of examinees who answer the question wrong.

<table>
<thead>
<tr>
<th>TestId</th>
<th>ExamineeId</th>
<th>QuestionId</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>U₁</td>
<td>Q₁</td>
<td>1</td>
</tr>
<tr>
<td>T₁</td>
<td>U₁</td>
<td>Q₂</td>
<td>1</td>
</tr>
<tr>
<td>T₁</td>
<td>U₂</td>
<td>Q₁</td>
<td>0</td>
</tr>
<tr>
<td>T₁</td>
<td>U₂</td>
<td>Q₂</td>
<td>1</td>
</tr>
<tr>
<td>T₂</td>
<td>U₄</td>
<td>Q₃</td>
<td>1</td>
</tr>
<tr>
<td>T₂</td>
<td>U₅</td>
<td>Q₃</td>
<td>1</td>
</tr>
<tr>
<td>T₂</td>
<td>U₆</td>
<td>Q₃</td>
<td>0</td>
</tr>
</tbody>
</table>

(T1) Q1: \((1+0)/2=0.5\)  
(T1) Q2: 0  
(T2) Q3: 0.33
Background

- Traditional solutions resort to expertise
  - Experts Labeling
    - Subjective
    - Biases on different experts, thus sometimes misleading
  - Artificial test organization
    - Labor intensive
    - Confidentiality
- Human-based solutions cannot applied to large-scale Question Difficulty Prediction (QDP)
Research Problem

- Urgent issue: **Question Difficulty Prediction (QDP)**
  - How to automatically predict question difficulty without manual intervention?
- Opportunity
  - Historical test logs of examinees
  - Text materials of questions
- This paper focuses on English Reading Problems
Challenge 1 for QDP

- Requires an **unified way** to understand and represent them from a semantic perspective.
  - Multiple parts of question texts
    - Document (TD)
    - Question (TQ)
    - Options (TO)
It is necessary and hard to **distinguish the importance** of text materials to a specific question

- Different questions concern different parts of texts
  - Q1 concentrates more on the highlighted “blue”
  - Q2 focuses more on the “green”

(D) Larry was on another of his underwater expeditions but this time, it was different. He decided to take his daughter along with him. She was only ten years old. [...] Dangerous areas did not prevent him from continuing his search. Sometimes, he was limited to a cage underwater but that did not bother him. [...] Already, she looked like she was much braver than had been then. This was the key to a successful underwater expedition.

(TQ) Q1: In what way was this expedition different for Larry?
A. His daughter had grown up.
B. He had become a famous diver.
C. His father would dive with him.
D. His daughter would dive with him.

(TQ) Q2: Why did Larry have to stay in a cage underwater sometimes?
A. To protect himself from danger.
B. To dive into the deep water.
C. To admire the underwater view.
D. To take photo more conveniently.
Challenge 3 for QDP

- It is necessary to take these **difficulty biases** into consideration for question difficulty prediction
  - Different questions are **incomparable** in different tests
    - Q2 with difficulty 0.6 in T1
    - Q1 with difficulty 0.37 in T2
Related Work for QDP

- **Education Psychology**
  - Possible factors contributed to question difficulty
    - Question attributes, i.e., question types (structures)
    - Examinee knowledge mastering degree
  - Cognitive Diagnosis Assessment (CDA)
    - Question difficulty obtained from examinees’ responses

- **Nature Language Process**
  - Understanding and representations of all text materials
    - Question selection
    - Textual entailment
    - Machine comprehension

- Machine abilities V.S. Question difficulty e.g., word reasoning

Take a lot of human effort. Not an automatic solution.
<table>
<thead>
<tr>
<th></th>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Background and Related Work</td>
</tr>
<tr>
<td>2</td>
<td>Problem Definition</td>
</tr>
<tr>
<td>3</td>
<td>Study Overview</td>
</tr>
<tr>
<td>4</td>
<td>TACNN Framework</td>
</tr>
<tr>
<td>5</td>
<td>Experiments</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion and Future Work</td>
</tr>
</tbody>
</table>
Problem Definition

- **Given**: questions of READING problems with corresponding text materials
- **Given**: historical examinees’ test logs.
- **Goal**: Automatically predict question difficulty in newly-conduct tests

Table 2: Examples of question instances combined with test logs and question materials.

<table>
<thead>
<tr>
<th>Difficulty (P)</th>
<th>QuestionId (Q)</th>
<th>TestId (T)</th>
<th>Document (TD)</th>
<th>Question (TQ)</th>
<th>Options (TO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4276</td>
<td>$Q_1$</td>
<td>$T_1$</td>
<td>Larry was on...</td>
<td>In what way...</td>
<td>His daughter had...</td>
</tr>
<tr>
<td>0.4827</td>
<td>$Q_2$</td>
<td>$T_1$</td>
<td>Larry was on...</td>
<td>Why did Larry...</td>
<td>To protect himself...</td>
</tr>
<tr>
<td>0.5494</td>
<td>$Q_3$</td>
<td>$T_1$</td>
<td>Larry was on...</td>
<td>What can be...</td>
<td>Larry had some...</td>
</tr>
<tr>
<td>?</td>
<td>$Q_4$</td>
<td>$T_2$</td>
<td>Are you...</td>
<td>Why do people...</td>
<td>They eat too...</td>
</tr>
</tbody>
</table>

Table 1: A toy example of test logs.

<table>
<thead>
<tr>
<th>TestId</th>
<th>ExamineeId</th>
<th>QuestionId</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>$U_1$</td>
<td>$Q_1$</td>
<td>1</td>
</tr>
<tr>
<td>$T_1$</td>
<td>$U_1$</td>
<td>$Q_2$</td>
<td>1</td>
</tr>
<tr>
<td>$T_1$</td>
<td>$U_2$</td>
<td>$Q_1$</td>
<td>0</td>
</tr>
<tr>
<td>$T_1$</td>
<td>$U_2$</td>
<td>$Q_2$</td>
<td>1</td>
</tr>
<tr>
<td>$T_2$</td>
<td>$U_4$</td>
<td>$Q_3$</td>
<td>1</td>
</tr>
<tr>
<td>$T_2$</td>
<td>$U_5$</td>
<td>$Q_3$</td>
<td>1</td>
</tr>
<tr>
<td>$T_2$</td>
<td>$U_6$</td>
<td>$Q_3$</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) A READING problem
Outline

1. Background and Related Work
2. Problem Definition
3. Study Overview
4. TACNN Framework
5. Experiments
6. Conclusion and Future Work
Study Overview

- Two-stage solution
  - Training stage
    - TACNN
    - Training strategy
  - Testing stage
    - Predict difficulty
# Outline

<table>
<thead>
<tr>
<th></th>
<th>Background and Related Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Problem Definition</td>
</tr>
<tr>
<td>3</td>
<td>Study Overview</td>
</tr>
<tr>
<td>4</td>
<td>TACNN Framework</td>
</tr>
<tr>
<td>5</td>
<td>Experiments</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion and Future Work</td>
</tr>
</tbody>
</table>
TACNN Framework

- Test-dependent Attention-based Convolutional Neural Network (TACNN)
  - Learning all text materials of each question from a sentence semantic perspective
    — CNN-based architecture
  - Learns attention representations for each question by qualifying the contributions of its text materials
    — Attention strategy
  - Wipe out the difficulty biases in different tests for training
    — Test-dependent strategy

Challenge 1: unified way
Challenge 2: qualify contributions
Challenge 3: Difficulty biases
TACNN Framework

- Four Layers

Figure 3: TACNN framework. The numbers in TACNN are the dimensions of corresponding feature vectors.
TACNN Framework – Input

- **Goal**: learn sentence representations from word perspective
- For each question (Text materials)
  - Document (TD)
    - Sequence sentences
  - Question (TQ)
    - One sentence
  - Options (TO)
    - Four sentences
- For each sentence
  - Sequence words
- For each word
  - Embedding
TACNN Framework – Sentence CNN

- **Goal:** learn sentence representations from semantic perspective

- CNN-based architecture
  - Capture dominated information
    - = Reading habit
  - Learn deep comparable semantic representations
  - Reduces the model complexity
TACNN Framework – Sentence CNN

- A variant of traditional CNN
  - Four Convolution (3 wide + 1 narrow)
  - Four pooling

\[ \overrightarrow{h}^c_i = \sigma(G \cdot [w_{i-k+1} \oplus \ldots \oplus w_i] + b), \]

\[ \overrightarrow{h}^{cp}_i = \left[ \max \left[ \overrightarrow{h}^c_{i-p+1,1} \right], \ldots, \max \left[ \overrightarrow{h}^c_{i-p+1,d} \right] \right] \]

Figure 4: Sentence CNN, which contains several layers of convolution and p-max pooling.
TACNN Framework - Attention Layer

- **Goal:**
  - Qualify the contributions of text materials to a specific question
  - Learn the attention representations

- Considering both documents and options level

\[
DA_i = \sum_{j=1}^{M} \alpha_j s_j^T D_i, \quad \alpha_j = \cos(s_j^T D_i, s_j^T Q_i),
\]

- **Attention vector**
- **Attention score**
TACNN Framework – Predict Layer

- **Goal:** predicting question difficulty
  - Document attention vector
  - Option attention vector
  - Question vector

\[ o_i = ReLU(W_1 \cdot [DA_i \oplus OA_i \oplus s^{TQ_i}] + b_1), \]
\[ \tilde{P}_i = Sigmoid(W_2 \cdot o_i + b_2), \]
TACNN — training strategy

- How to train?
- Supervised way: leverage historical test logs of examinees

Figure 3: TACNN framework. The numbers in TACNN are the dimensions of corresponding feature vectors.
Biases: question difficulties are test-dependent

- Different questions in different tests are incomparable, i.e., Q1 and Q3
- Different questions in same tests are comparable, i.e., Q1 and Q2

<table>
<thead>
<tr>
<th>TestId</th>
<th>ExamineeId</th>
<th>QuestionId</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>U₁</td>
<td>Q₁</td>
<td>1</td>
</tr>
<tr>
<td>T₁</td>
<td>U₁</td>
<td>Q₂</td>
<td>1</td>
</tr>
<tr>
<td>T₁</td>
<td>U₂</td>
<td>Q₁</td>
<td>0</td>
</tr>
<tr>
<td>T₁</td>
<td>U₂</td>
<td>Q₂</td>
<td>1</td>
</tr>
<tr>
<td>T₂</td>
<td>U₄</td>
<td>Q₃</td>
<td>1</td>
</tr>
<tr>
<td>T₂</td>
<td>U₅</td>
<td>Q₃</td>
<td>1</td>
</tr>
<tr>
<td>T₂</td>
<td>U₆</td>
<td>Q₃</td>
<td>0</td>
</tr>
</tbody>
</table>

(T1) Q1: \(\frac{1+0}{2}=0.5\)
(T1) Q2: 0
(T2) Q3: 0.33

Which is more difficult?
TACNN — training strategy

- Test-dependent pairwise training objective
  - Training “gap” from two question difficulties

\[
J(\Theta) = \sum_{(T_t, Q_i, Q_j)} \left( (P_i^t - P_j^t) - (M(Q_i) - M(Q_j)) \right)^2 + \lambda \Theta ||\Theta_M||^2, \quad (6)
\]

- Minimize the objective function by AdaDelta

Qi, Qj in same test Tt
Prediction of Qj
Prediction of Qi
TACNN — testing stage

- After training, we can predict question difficulty from text perspectives, e.g., words or sentences
- More application
  - Automatically label question for large-scale systems
  - Help decide whether the question to choose into the test paper or not.
Experiments

- Experiments dataset
  - Supplied by IFLYTEK
  - Collected from real-world standard tests for READING problems in Chinese senior high schools from the year 2014 to 2016

Table 3: The statistics of the dataset.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td># of test logs</td>
<td>28,818,047</td>
</tr>
<tr>
<td># of examinees</td>
<td>1,019,415</td>
</tr>
<tr>
<td># of tests</td>
<td>4,085</td>
</tr>
<tr>
<td># of READINGs</td>
<td>8,220</td>
</tr>
<tr>
<td># of questions</td>
<td>30,817</td>
</tr>
<tr>
<td>Average questions per test</td>
<td>14.167</td>
</tr>
<tr>
<td>Average tests per question</td>
<td>1.877</td>
</tr>
</tbody>
</table>

Figure 5: Statistics of observed records.

(a) Sentences distribution
(b) Words distribution
Experiments

- **Baseline methods**
  - Variants of TACNN: CNN, ACNN, TCNN
    - To validate the performance of each component in TACNN
  - Machine comprehension (MC) model: HABCNN
    - The most similar network architecture to ours

- **Evaluation metrics**
  - RMSE
  - DOA: Measure the percentage of correctly ranked difficulties of question pairs
  - PCC: Pearson Correlation Coefficient
  - PR: the percentage of tests which pass t-test at confidence level of 0.05
Experiments

- Overall results

- Attention strategy and test-dependent training strategy do effectively

- Solutions to MC task is unsuitable for QDP

- Demonstrates the rationality of pairwise training strategy
Experiments

- Experts comparisons

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.41</td>
<td>0.21</td>
<td>0.18</td>
<td>0.13</td>
<td>0.38</td>
<td>-0.08</td>
<td>-0.04</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>T2</td>
<td>0.63</td>
<td>0.68</td>
<td>0.45</td>
<td>0.32</td>
<td>0.52</td>
<td>-0.01</td>
<td>-0.44</td>
<td>0.53</td>
<td>0.37</td>
</tr>
<tr>
<td>T3</td>
<td>0.78</td>
<td>0.70</td>
<td>0.52</td>
<td>0.63</td>
<td>0.28</td>
<td>0.44</td>
<td>-0.29</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>T4</td>
<td>0.63</td>
<td>0.40</td>
<td>-0.09</td>
<td>0.07</td>
<td>0.31</td>
<td>0.48</td>
<td>-0.40</td>
<td>0.58</td>
<td>-0.08</td>
</tr>
<tr>
<td>T5</td>
<td>0.53</td>
<td>0.56</td>
<td>0.39</td>
<td>0.32</td>
<td>0.29</td>
<td>0.29</td>
<td>0.43</td>
<td>0.51</td>
<td>0.47</td>
</tr>
<tr>
<td>T6</td>
<td>0.47</td>
<td>0.22</td>
<td>0.21</td>
<td>0.01</td>
<td>0.27</td>
<td>-0.23</td>
<td>0.10</td>
<td>0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>T7</td>
<td>0.81</td>
<td>0.73</td>
<td>0.58</td>
<td>0.29</td>
<td>0.72</td>
<td>0.72</td>
<td>0.70</td>
<td>0.59</td>
<td>0.69</td>
</tr>
<tr>
<td>T8</td>
<td>0.77</td>
<td>0.45</td>
<td>0.35</td>
<td>0.45</td>
<td>0.24</td>
<td>0.14</td>
<td>0.19</td>
<td>0.45</td>
<td>0.64</td>
</tr>
<tr>
<td>T9</td>
<td>0.81</td>
<td>0.55</td>
<td>0.25</td>
<td>0.54</td>
<td>0.35</td>
<td>0.53</td>
<td>0.13</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>T10</td>
<td>0.76</td>
<td>0.57</td>
<td>0.49</td>
<td>-0.13</td>
<td>0.72</td>
<td>0.25</td>
<td>0.22</td>
<td>0.32</td>
<td>0.60</td>
</tr>
<tr>
<td>T11</td>
<td>0.90</td>
<td>0.77</td>
<td>0.44</td>
<td>0.57</td>
<td>0.59</td>
<td>0.41</td>
<td>0.36</td>
<td>0.08</td>
<td>0.83</td>
</tr>
<tr>
<td>T12</td>
<td>0.60</td>
<td>0.62</td>
<td>0.59</td>
<td>0.73</td>
<td>0.60</td>
<td>0.54</td>
<td>0.48</td>
<td>0.62</td>
<td>0.54</td>
</tr>
<tr>
<td>Avg</td>
<td>0.68</td>
<td>0.54</td>
<td>0.36</td>
<td>0.33</td>
<td>0.44</td>
<td>0.29</td>
<td>0.12</td>
<td>0.39</td>
<td>0.44</td>
</tr>
<tr>
<td>Std</td>
<td>0.14</td>
<td>0.18</td>
<td>0.19</td>
<td>0.26</td>
<td>0.17</td>
<td>0.27</td>
<td>0.34</td>
<td>0.19</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- Predictions from experts are **not always consistent**
- Expert predictions are **subjective**, which are hardly of the same mind.
- Expert predictions may sometimes misleading
Experiments

- Model explanatory power (model visualization)
  - Document-level (Q1)

Good way for a question to capture key information for model explanations
Conclusion

- Proposed an **unified TACNN framework** for question difficulty prediction task.

- TACNN integrated **two critical components**, i.e., Sentence CNN Layer and Attention Layer, which can exactly **learn question representations** for reading problems from semantic perspective.

- Proposed a **test-dependent pairwise strategy** for training TACNN and generating the difficulty prediction values.

- Experiments on real-world dataset demonstrated both the **effectiveness** and **explanatory power** of TACNN.
Future Work

- We will make our efforts to design a more efficient learning algorithm for TACNN.

- We are also willing to extend TACNN to solve QDP task in:
  - Other types of problems in English tests, e.g., LISTENING, WRITING
  - Other subjects, e.g., MATH
Thanks!