

Question Difficulty Prediction for READING Problems in Standard Tests

Zhenya Huang[†], Qi Liu^{†*}, Enhong Chen[†], Hongke Zhao[†], Mingyong Gao[‡], Si Wei[‡], Yu Su^ᵇ, 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

^ᵇSchool of Computer Science and Technology, Anhui University, yusu@iflytek.com

Reporter: Zhenya Huang

Date: Feb. 7th, 2017

Outline

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

Background

- 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 1: A toy example of test logs.

TestId	ExamineeId	QuestionId	Score
T_1	U_1	Q_1	1
T_1	U_1	Q_2	1
T_1	U_2	Q_1	0
T_1	U_2	Q_2	1
T_2	U_4	Q_3	1
T_2	U_5	Q_3	1
T_2	U_6	Q_3	0
...

(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

The image shows a screenshot of a Chinese exam score report. On the left, there's a smaller table with columns for subject, score, and date. On the right, a larger table titled '七、八、九、十、十一、十二' (Grades 7-12) shows scores for various subjects. A red box highlights a row in the larger table, and a red arrow points to it from the smaller table. The score '81' is prominently displayed in a red box.

(TD) 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?

(TD) {
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?

(TD) {
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 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)

(TD) 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?

(TO) {
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?

(TO) {
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 2 for QDP

- 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”

(TD) 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?

- (TO) {
- 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.

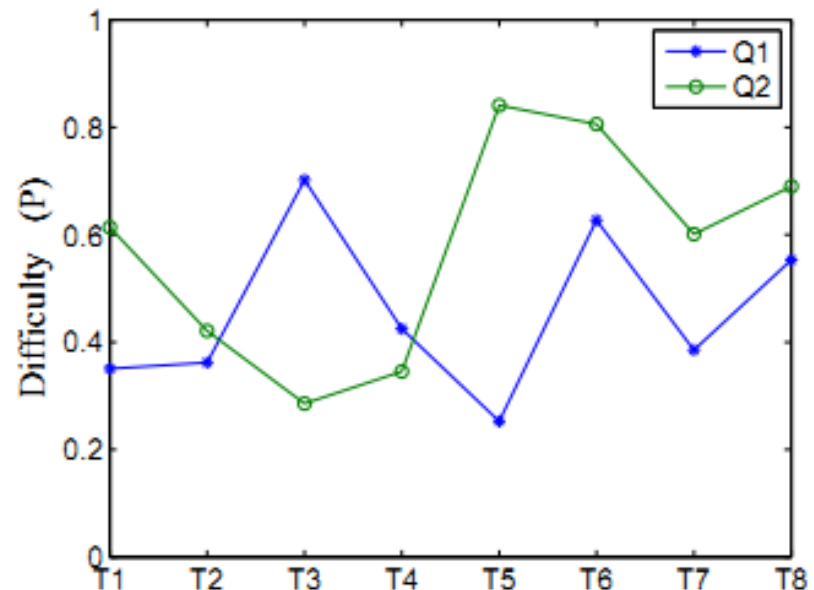
(TQ)

Q2: Why did Larry have to stay in a cage underwater sometimes?

- (TO) {
- 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

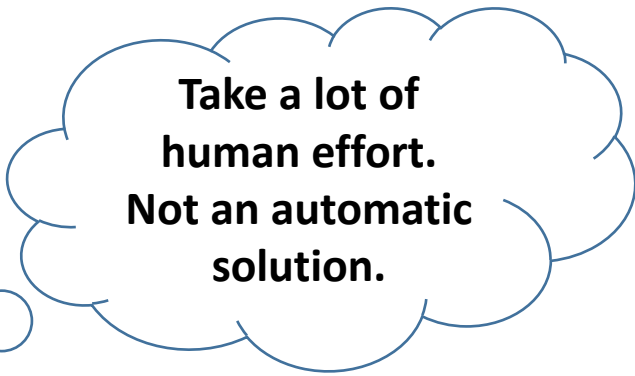
- It is necessary to take these **difficulty biases** into consideration for question difficulty prediction
 - Different questions are **incomparable** in different tests
 - Q2 with difficulty 0.6 in T1
 - Q1 with difficulty 0.37 in T2



Related Work for QDP

➤ Education Psychology

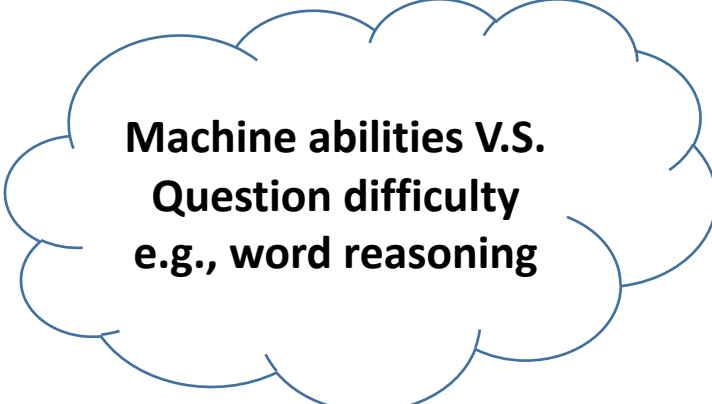
- Possible factors contributed to question difficulty
 - Question attributes, i.e., question types (structures)
 - Examinee knowledge mastering degree
- Cognitive Diagnosis Assessment (CDA)
 - Question difficulty obtained from examinees' responses



**Take a lot of
human effort.
Not an automatic
solution.**

➤ 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**

Outline

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

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.

Difficulty (P)	QuestionId (Q)	TestId (T)	Text Materials						
			Document (TD)	Question (TQ)	Options (TO)				
0.4276	Q_1	T_1	Larry was on...	In what way...	His daughter had...	He had become...	His father...	His daughter...	
0.4827	Q_2	T_1	Larry was on...	Why did Larry...	To protect himself...	To dive into...	To admire the...	To take photo...	
0.5494	Q_3	T_1	Larry was on...	What can be...	Larry had some...	Larry liked the...	Divers had to...	Ten-year-old...	
?	Q_4	T_2	Are you...	Why do people...	They eat too...	They sleep too...	Their body...	The weather...	

Table 1: A toy example of test logs.

TestId	ExamineeId	QuestionId	Score
T_1	U_1	Q_1	1
T_1	U_1	Q_2	1
T_1	U_2	Q_1	0
T_1	U_2	Q_2	1
T_2	U_4	Q_3	1
T_2	U_5	Q_3	1
T_2	U_6	Q_3	0
...

(TD)	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?
(TO)	<ul style="list-style-type: none"> 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?
(TO)	<ul style="list-style-type: none"> 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.

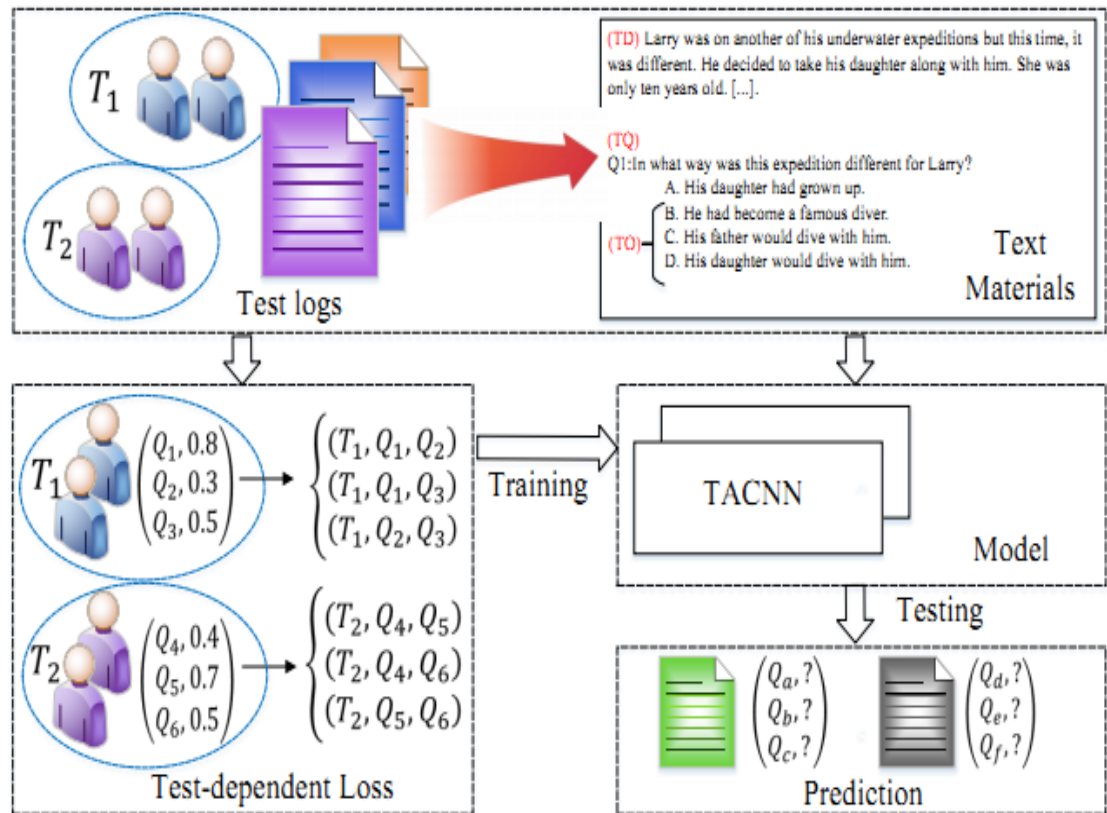
(a) A READING problem

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

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

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



**Challenge 1:
unified way**

- Learns **attention representations** for each question by qualifying the contributions of its text materials

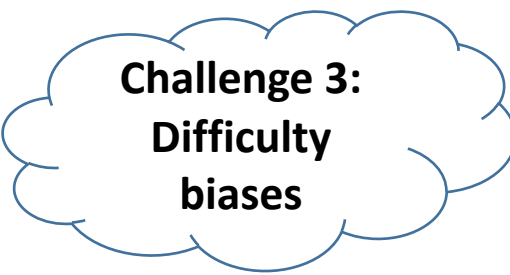
— Attention strategy



**Challenge 2:
qualify
contributions**

- Wipe out the **difficulty biases** in different tests for training

— Test-dependent strategy



**Challenge 3:
Difficulty
biases**

TACNN Framework

➤ Four Layers

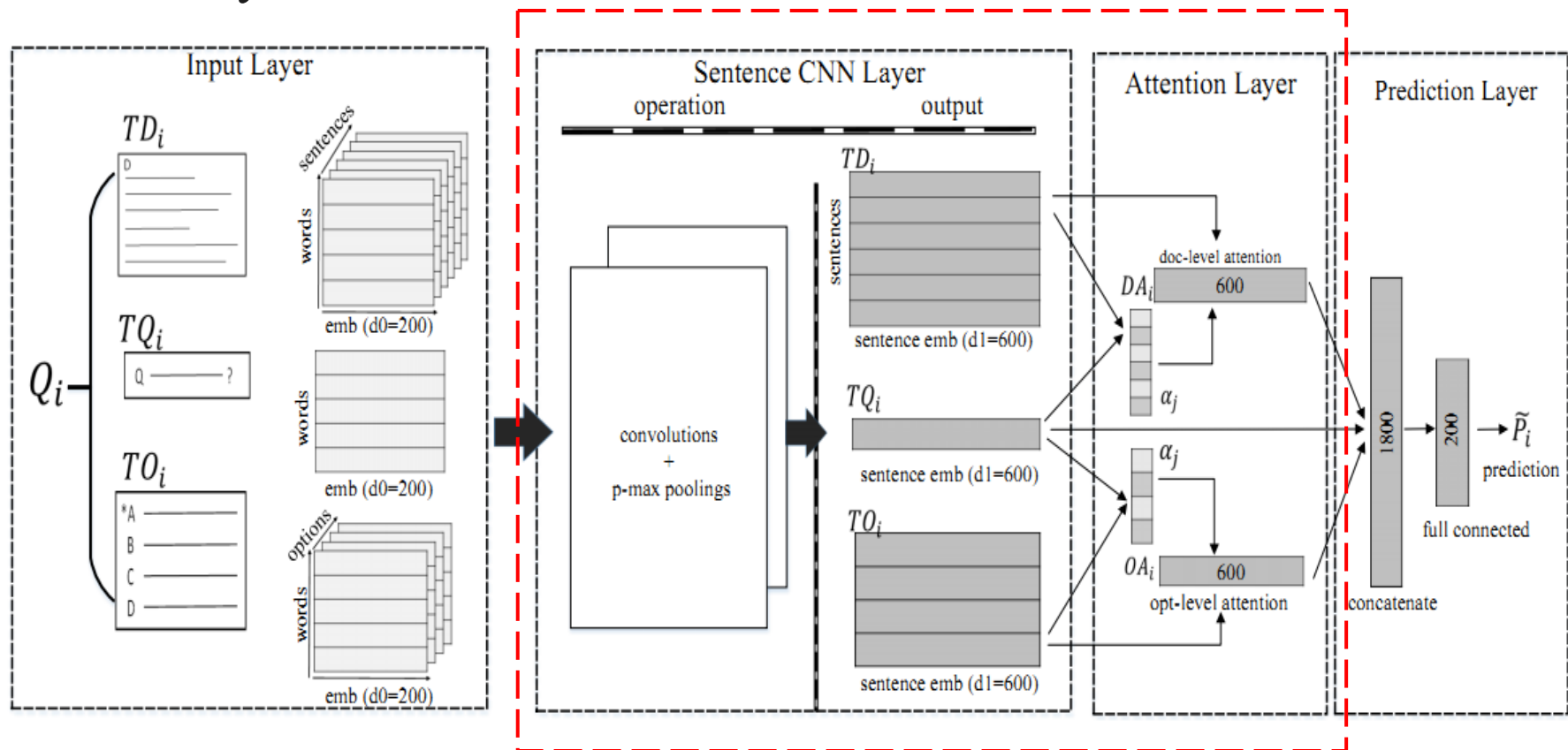
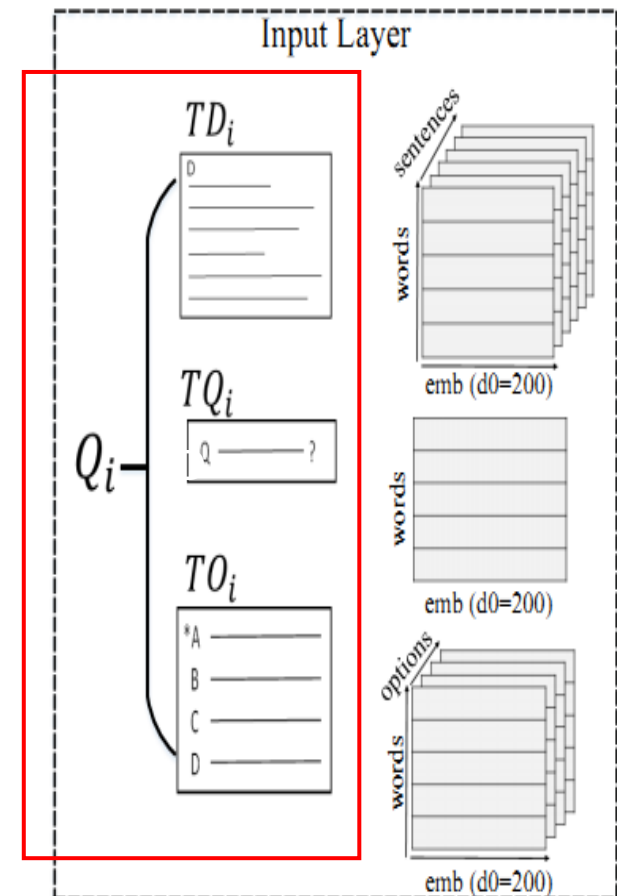


Figure 3: TACNN framework. The numbers in TACNN are the dimensions of corresponding feature vectors.

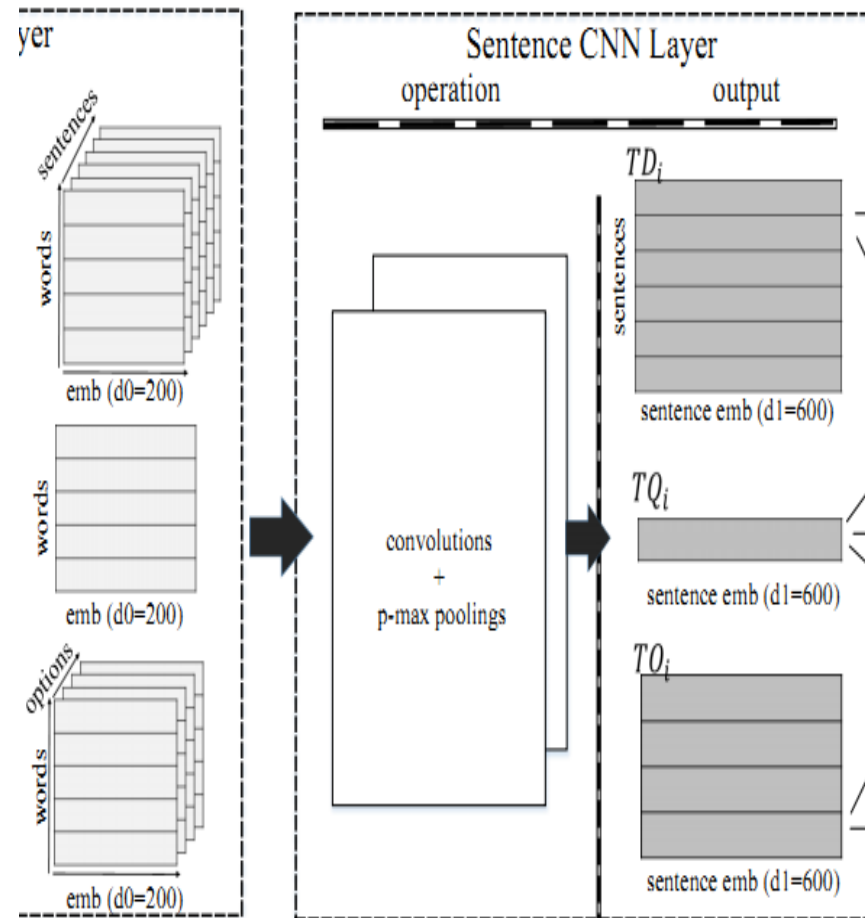
TACNN Framework – Input

- **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



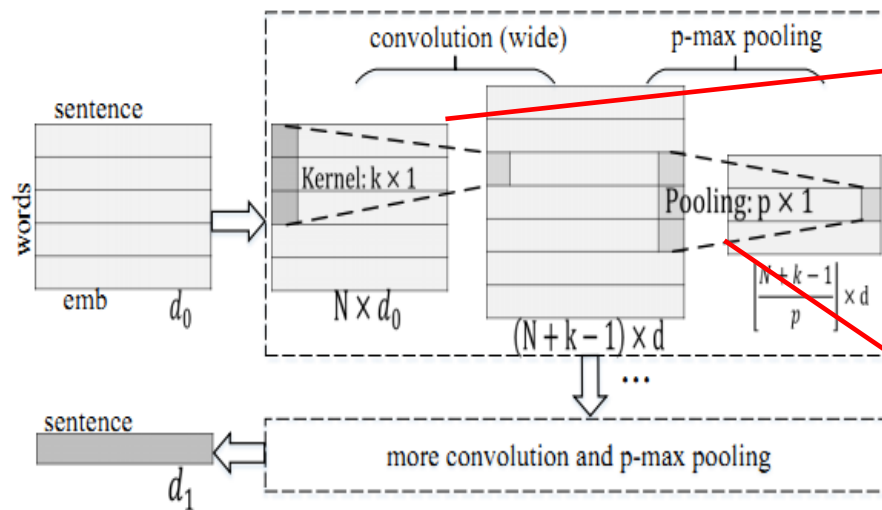
TACNN Framework – Sentence CNN

- **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



$$\vec{h}_i^c = \sigma(\mathbf{G} \cdot [w_{i-k+1} \oplus \dots \oplus w_i] + \mathbf{b}),$$

$$\vec{h}_i^{cp} = \left[\max \begin{bmatrix} h_{i-p+1,1}^c \\ \vdots \\ h_{i,1}^c \end{bmatrix}, \dots, \max \begin{bmatrix} h_{i-p+1,d}^c \\ \vdots \\ h_{i,d}^c \end{bmatrix} \right]$$

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

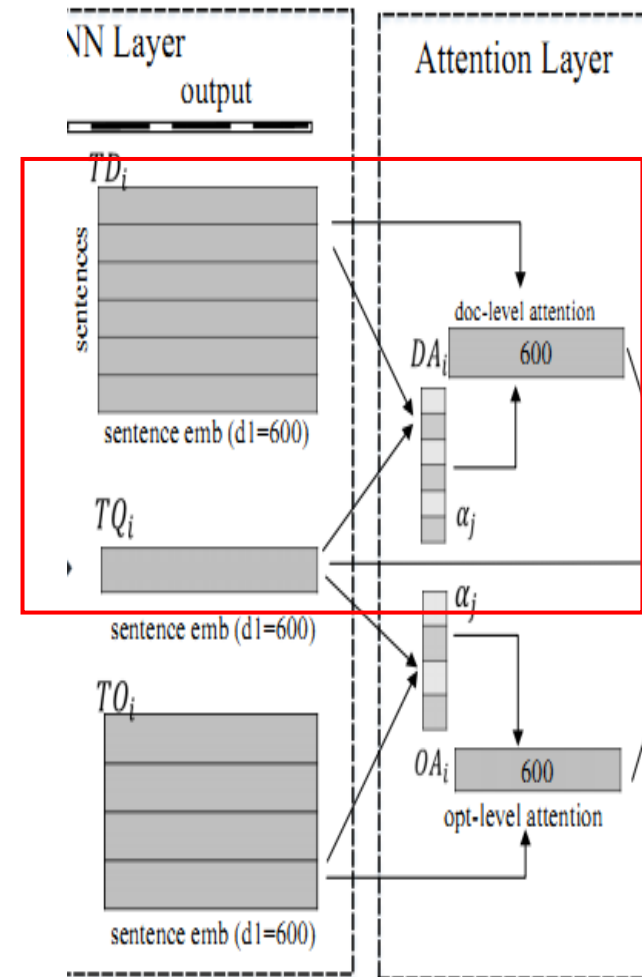
TACNN Framework - Attention Layer

- **Goal:**
 - Qualify the **contributions** of text materials to a specific question
 - Learn the **attention representations**
- Considering both **documents** and options level

$$DA_i = \sum_{j=1}^M \alpha_j s_j^{TD_i}, \quad \alpha_j = \cos(s_j^{TD_i}, s_j^{TQ_i}),$$

Attention
vector

Attention
score



TACNN Framework – Predict Layer

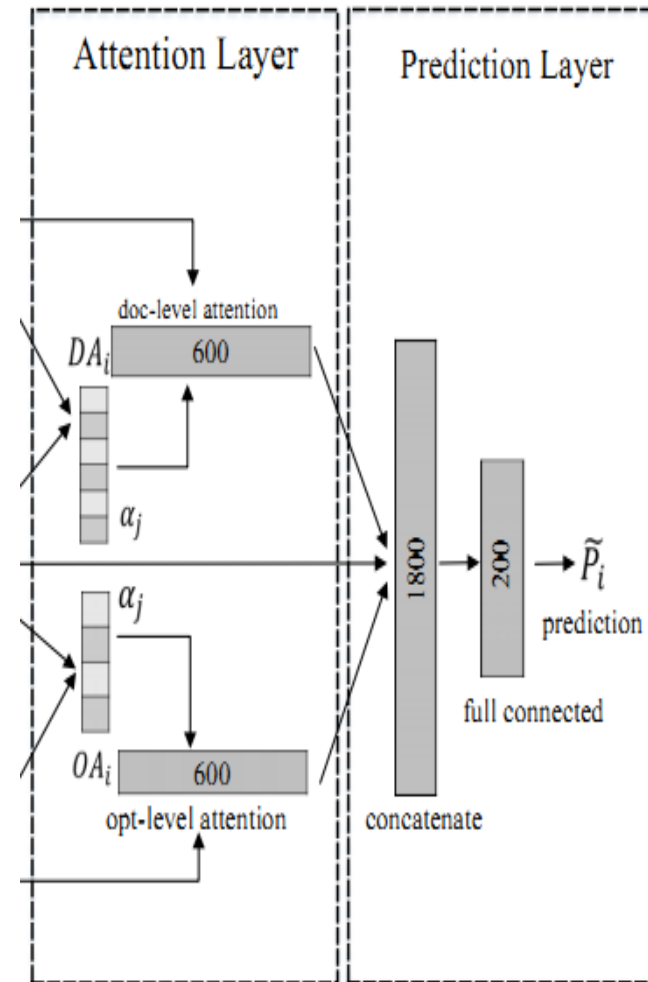
- **Goal:** predicting question difficulty
 - Document attention vector
 - Option attention vector
 - Question vector

Document attention vector

Question vector

$$o_i = \text{ReLU}(\mathbf{W}_1 \cdot [DA_i \oplus OA_i \oplus s^{TQ_i}] + \mathbf{b}_1),$$
$$\tilde{P}_i = \text{Sigmoid}(\mathbf{W}_2 \cdot o_i + \mathbf{b}_2),$$

Option attention vector



TACNN — training strategy

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

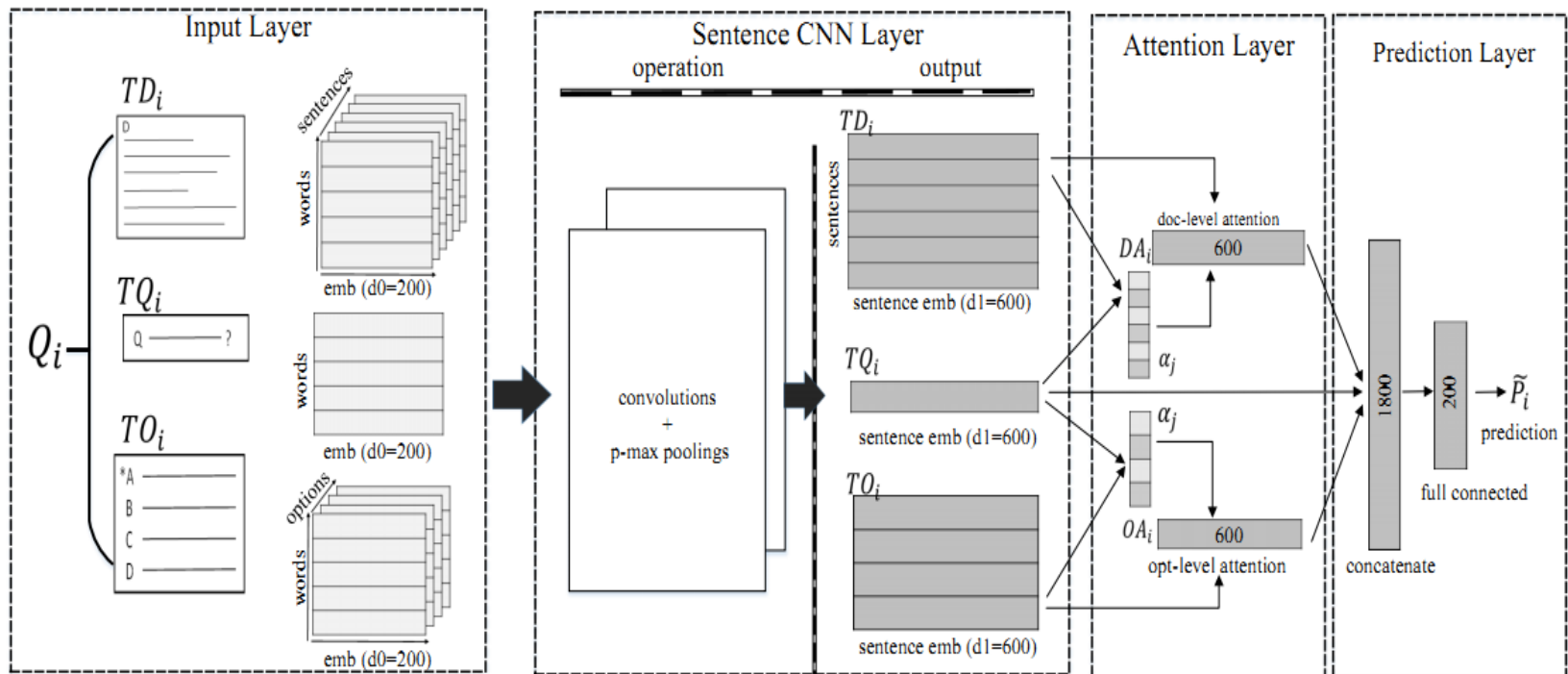


Figure 3: TACNN framework. The numbers in TACNN are the dimensions of corresponding feature vectors.

TACNN — training strategy

- 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 1: A toy example of test logs.

TestId	ExamineeId	QuestionId	Score
T_1	U_1	Q_1	1
T_1	U_1	Q_2	1
T_1	U_2	Q_1	0
T_1	U_2	Q_2	1
T_2	U_4	Q_3	1
T_2	U_5	Q_3	1
T_2	U_6	Q_3	0
...

(T1) Q1: $(1+0)/2=0.5$

(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

$$\mathcal{J}(\Theta) = \sum_{(T_t, Q_i, Q_j)} ((P_i^t - P_j^t) - (\mathcal{M}(Q_i) - \mathcal{M}(Q_j)))^2 + \lambda_{\Theta} \|\Theta_{\mathcal{M}}\|^2, \quad (6)$$

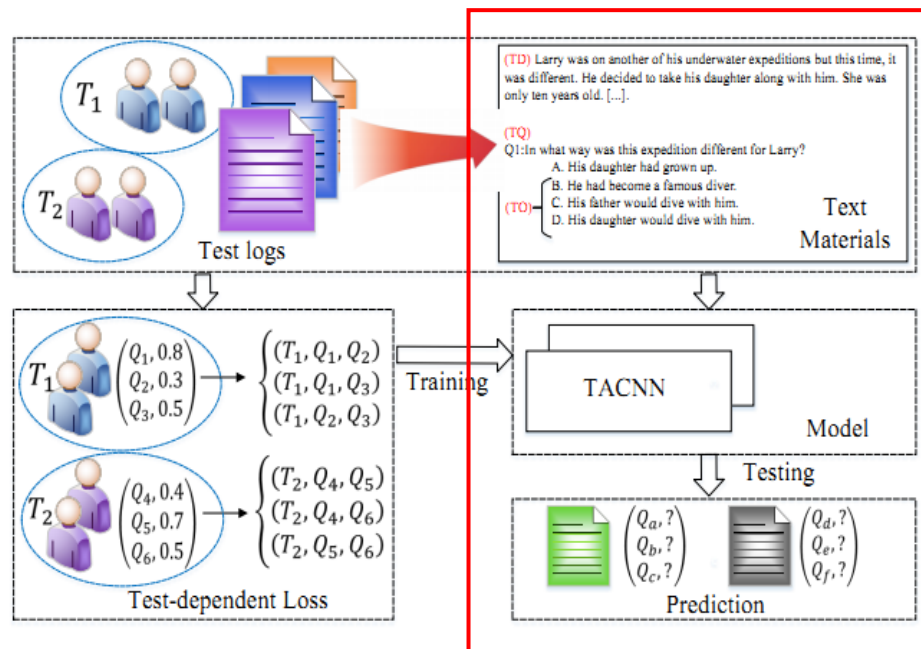
Diagram illustrating the components of the objective function $\mathcal{J}(\Theta)$:

- (T_t, Q_i, Q_j) : Qi, Qj in same test T_t
- $\mathcal{M}(Q_i)$: Prediction of Q_i
- $\mathcal{M}(Q_j)$: Prediction of Q_j

- Minimize the objective function by AdaDelta

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.



Outline

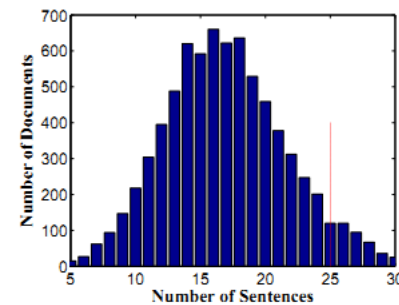
- 1 **Background and Related Work**
- 2 **Problem Definition**
- 3 **Study Overview**
- 4 **TACNN Framework**
- 5 Experiments**
- 6 **Conclusion and Future Work**

Experiments

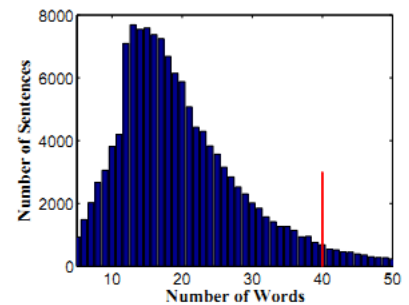
- Experiments dataset
 - Supplied by IFLYTEK
 - Collected from **real-world standard tests** for READING problems in Chinese senior high schools from the year 2014 to 2016

Table 3: The statistics of the dataset.

Statistics	Values
# of test logs	28,818,047
# of examinees	1,019,415
# of tests	4,085
# of READINGS	8,220
# of questions	30,817
Average questions per test	14.167
Average tests per question	1.877



(a) Sentences distribution



(b) Words distribution

Figure 5: Statistics of observed records.

Experiments

➤ Baseline methods

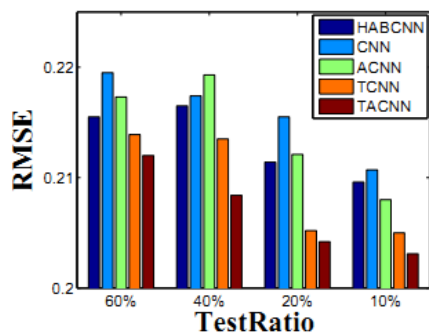
- Variants of TACNN: CNN, ACNN, TCNN
 - To validate the performance of **each component** in TACNN
- Machine comprehension (MC) model: HABCNN
 - The most similar **network architecture** to ours

➤ Evaluation metrics

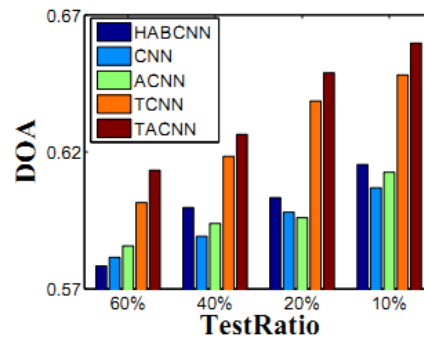
- RMSE
- DOA: Measure the percentage of correctly ranked difficulties of question pairs
- PCC: Pearson Correlation Coefficient
- PR: the percentage of tests which pass t-test at confidence level of 0.05

Experiments

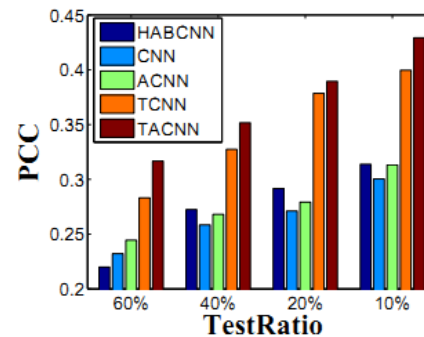
➤ Overall results



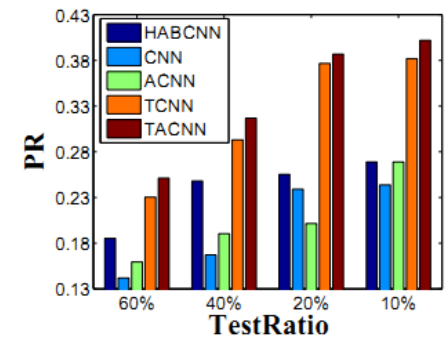
(a) The performance on RMSE



(b) The performance on DOA



(c) The performance on PCC



(d) The performance on PR

- 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 4: TACNN v.s. Experts on QDP task with PCC metric.

Test	TACNN	EpAvg	Ep1	Ep2	Ep3	Ep4	Ep5	Ep6	Ep7
T1	0.41	0.21	0.18	0.13	0.38	-0.08	-0.04	0.01	0.14
T2	0.63	0.68	0.45	0.32	0.52	-0.01	-0.44	0.53	0.37
T3	0.78	0.70	0.52	0.63	0.28	0.44	-0.29	0.45	0.52
T4	0.63	0.40	-0.09	0.07	0.31	0.48	-0.40	0.58	-0.08
T5	0.53	0.56	0.39	0.32	0.29	0.29	0.43	0.51	0.47
T6	0.47	0.22	0.21	0.01	0.27	-0.23	0.10	0.24	0.17
T7	0.81	0.73	0.58	0.29	0.72	0.72	0.70	0.59	0.69
T8	0.77	0.45	0.35	0.45	0.24	0.14	0.19	0.45	0.64
T9	0.81	0.55	0.25	0.54	0.35	0.53	0.13	0.32	0.36
T10	0.76	0.57	0.49	-0.13	0.72	0.25	0.22	0.32	0.60
T11	0.90	0.77	0.44	0.57	0.59	0.41	0.36	0.08	0.83
T12	0.60	0.62	0.59	0.73	0.60	0.54	0.48	0.62	0.54
Avg	0.68	0.54	0.36	0.33	0.44	0.29	0.12	0.39	0.44
Std	0.14	0.18	0.19	0.26	0.17	0.27	0.34	0.19	0.25

- 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)

(TD) 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?

(TO) {
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?

(TO) {
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.

Larry was on another of his underwater expeditions...
He decided to take his daughter along with him...
This would be her first trip with her father on what...
Larry first began diving when he was his daughter...
...
Then, there was the instructor.
He gave him a short lesson before allowing him...
...
After the first expedition, Larry is later diving...
There was never a dull moment. In his black and...
Dangerous areas did not prevent him from ...
Sometimes, he was limited to a cage underwater...
...
Larry has first expedition without his father was...
Fortunately for him, a man offered to take him...
...
He hoped she would be able to continue the...
This was the key to a successful underwater...

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

Outline

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

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

Q & A



Thanks!