Towards a Holistic Understanding of Mathematical Questions with Contrastive Pre-training

Yuting Ning,[1,2] **Zhenya Huang,**[1,2] **Xin Lin,**[1,2] **Enhong Chen,**[1,2*]
Shiwei Tong,[1,2] **Zheng Gong,**[1,2] **Shijin Wang**[2,3]

[1] Anhui Province Key Laboratory of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China
[2] State Key Laboratory of Cognitive Intelligence
[3] iFLYTEK AI Research (Central China), iFLYTEK Co., Ltd.
{ningyt, linx, tongsw, gz70229}@mail.ustc.edu.cn, {huangzhy, cheneh}@ustc.edu.cn, sjwang3@iflytek.com

Reporter: Yuting Ning

Outline

Background

- Online learning systems
  - collect massive educational questions
  - provide personalized applications based on the understanding of these questions
    - understand the difficulties of questions to recommend suitable questions for learners with different abilities
- Mathematical question understanding
  - more difficult to understand with special components (e.g., formulas) and complex mathematical logic
  - require more domain knowledge for comprehensive understanding
  - a crucial but challenging issue in intelligent education field

Background

- Task-specific Method
  - focus on one specific aspect of questions (e.g., difficulty)
  - learn from application tasks (e.g., difficulty estimation)
  - require massive expertise for human annotations and suffer from the label sparsity
- Pre-training Method
  - aim to learn comprehensive question representations on large-scale unlabeled data and benefit various applications
  - mainly focus on the details of question content
  - lack the consideration of holistic meanings from a mathematical perspective

Background

- Challenges
  - Compared with general texts and other educational questions, mathematical questions are more complex with special components (e.g., formulas), and require more mathematical knowledge and logic to understand.

formulas

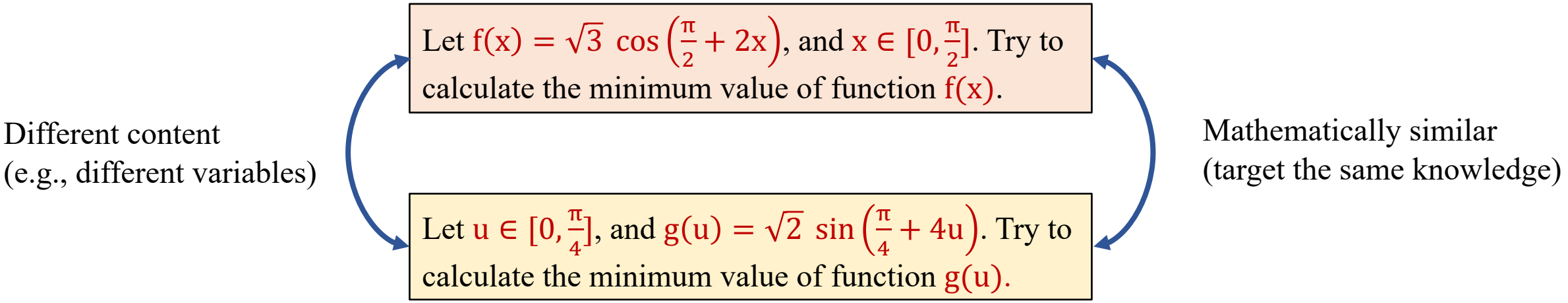
Let $f(x) = \sqrt{3}\ \cos\left(\frac{\pi}{2} + 2x\right)$, and $x \in [0, \frac{\pi}{2}]$. Try to calculate the minimum value of function $f(x)$.

Mathematical Question

Knowledge Concept:
Trigonometric Function

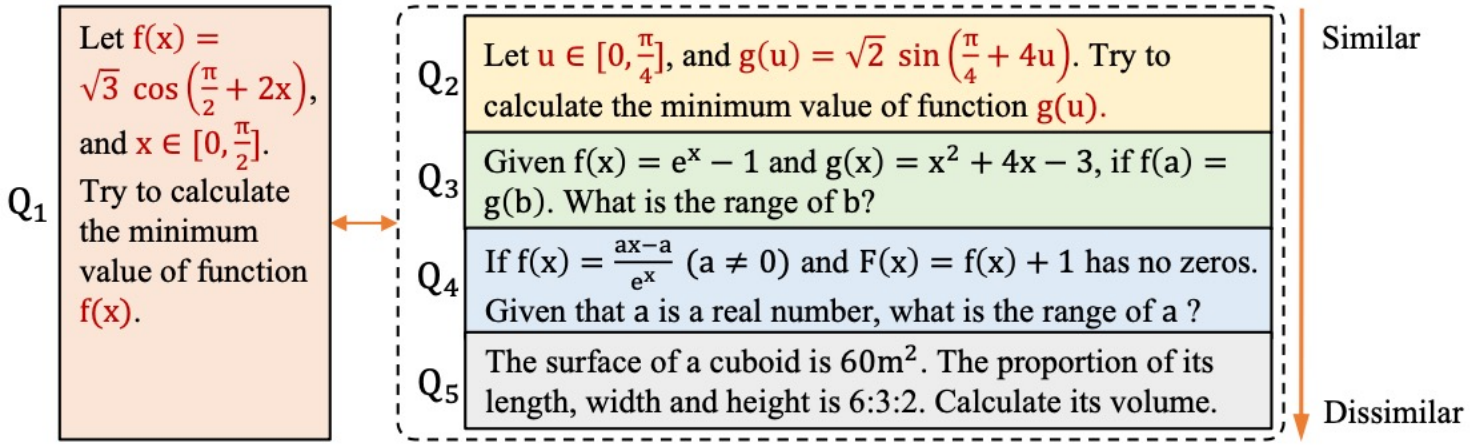
Background

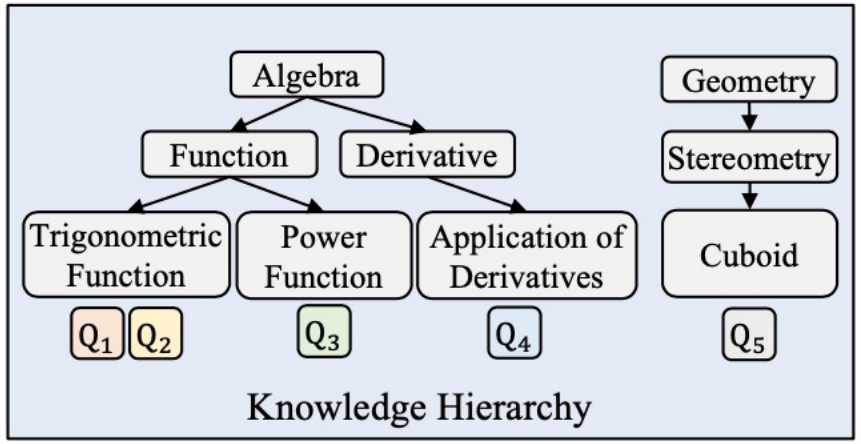
- Challenges
  - The holistic mathematical purposes of questions are more important than literal details.



Background

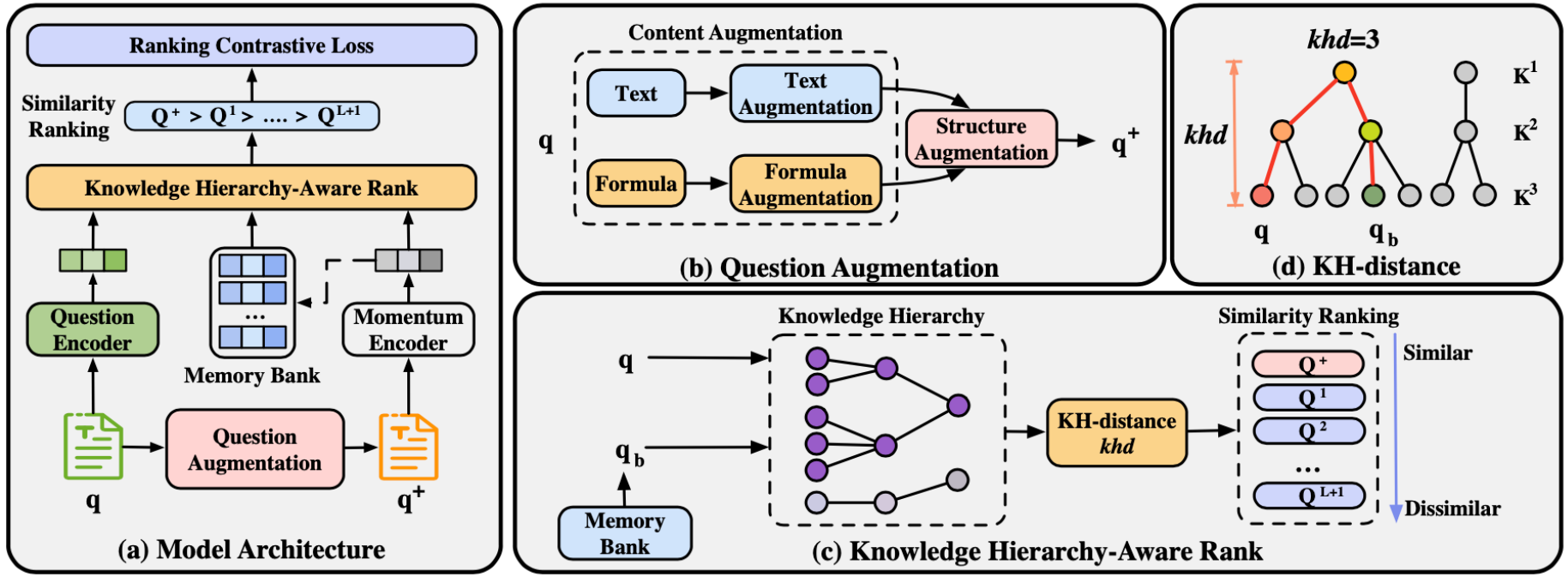
- Challenges
  - The related knowledge concepts play an important role in understanding mathematical questions, since they reflect the purposes and mathematical domain of questions.





Background

- Challenges
  - In this paper, we propose a novel contrastive pre-training method for holistically understanding mathematical questions (QuesCo).



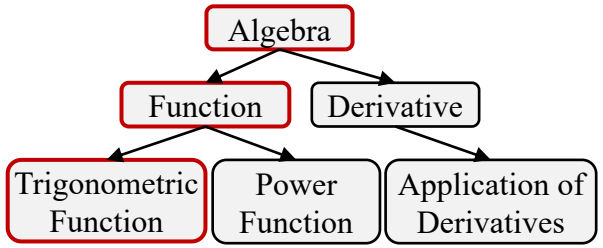
Outline

Problem Definition

- Mathematical Question
  - consist of content and related knowledge concepts $\boldsymbol{q} = (\boldsymbol{x}, \boldsymbol{k})$
  - content
    - a sequence of T tokens, denoted as $\boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$
    - each token x_i is either a text word or a formula symbol
  - related knowledge concepts
    - selected from a L-level knowledge hierarchy $KH = \{\mathcal{K}, \mathcal{E}\}$
    - $\boldsymbol{k} = \{k_1, k_2, \ldots, k_L\}$, where $k_i \in K_i$ and k_i is the ancestor of k_j $(j > i)$

Let f(x) $= \sqrt{3}\ \cos\left(\frac{\pi}{2} + 2x\right)$, and x $\in [0, \frac{\pi}{2}]$. Try to calculate the minimum value of function f(x).

Question Content



Related Knowledge Concepts

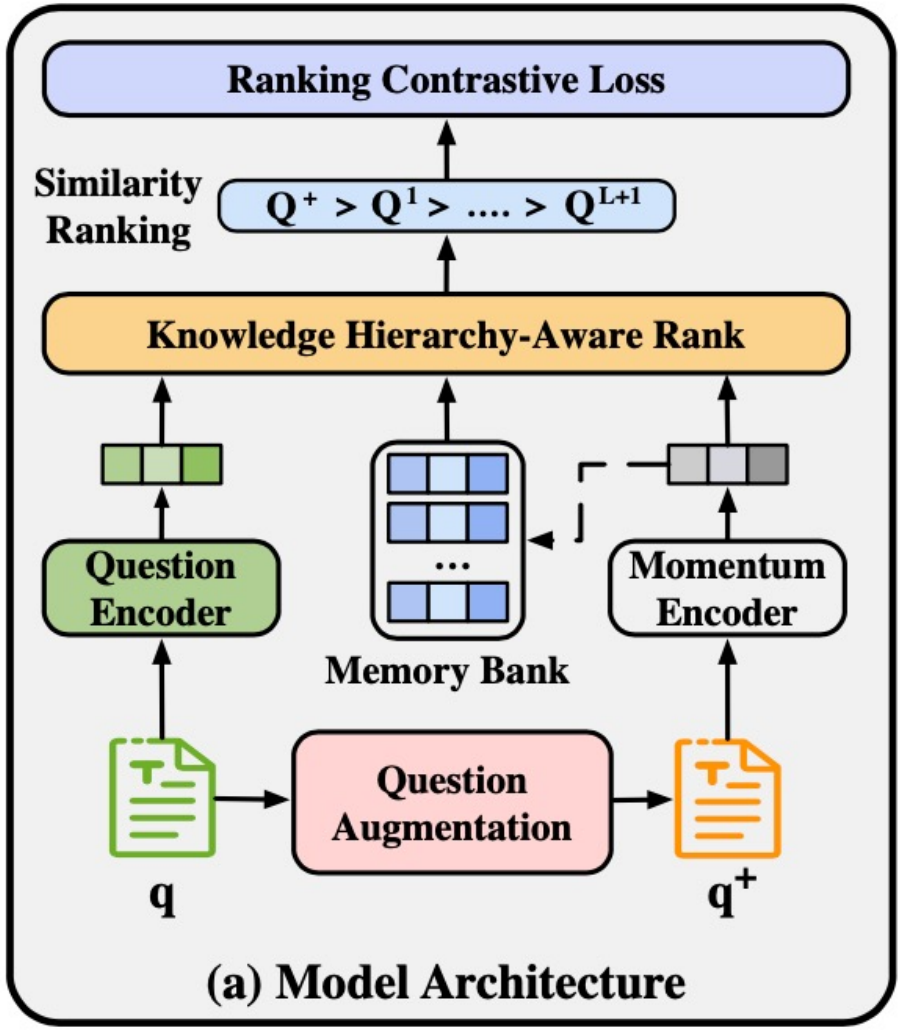
Problem Definition

- Question Representation Problem
  - Given
    - mathematical question $\boldsymbol{q} = (\boldsymbol{x}, \boldsymbol{k})$
  - Goal
    - represent $\boldsymbol{q}$ with a d-dimensional vector $\boldsymbol{v} \in \mathbb{R}^d$, which can be transferred to several downstream tasks and benefit their performances
    - the vector is expected
      - capture latent purposes of questions from a holistic perspective
      - contain the rich information in question content (e.g., formulas) and related knowledge concepts

Outline

QuesCo Framework

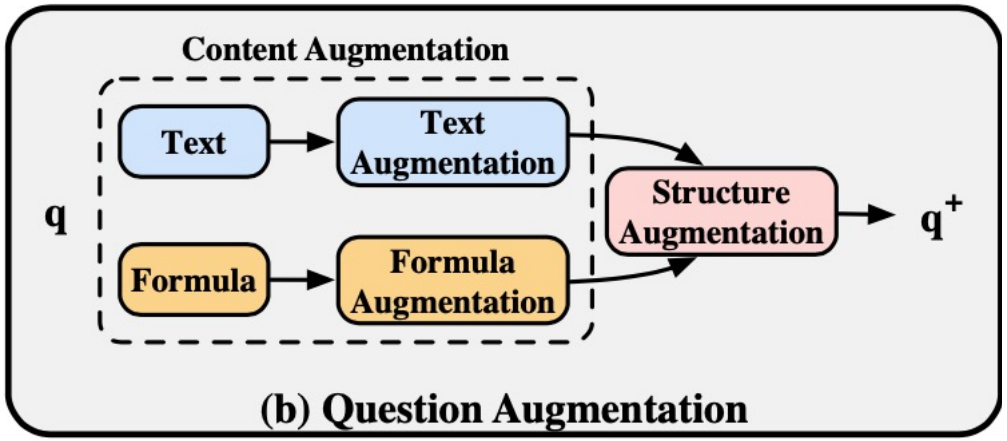
- QuesCo Framework
  - Based on contrastive learning, we propose a novel pre-training approach for mathematical questions
    - learn comprehensive question representations by pulling questions with more similar purposes closer than those with less similar purposes
  - How to construct contrastive pairs?
    - Question Augmentation
    - Knowledge Hierarchy-Aware Rank
  - How to optimize?
    - Ranking Contrastive Loss



QuesCo Framework

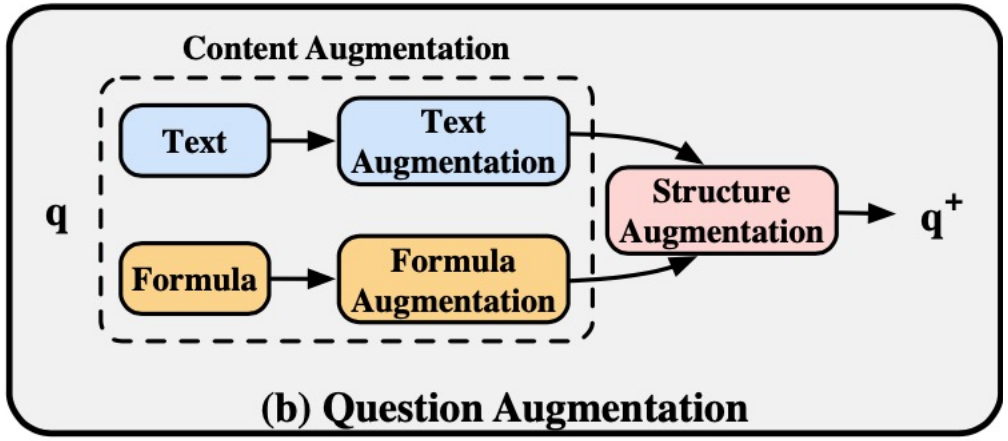
- Question Augmentation
  - **Goal:** To learn latent purposes of mathematical questions, generate mathematical questions with similar holistic purposes but diverse literal details
  - However, mathematical question content is complex with various components (i.e., plain text and formulas) and mathematical knowledge
  - Question structure also has unique logic (e.g., parallelism of conditional clauses)

- two-level augmentation strategies
  - Content-level
    - Text and formula
  - Structure-level



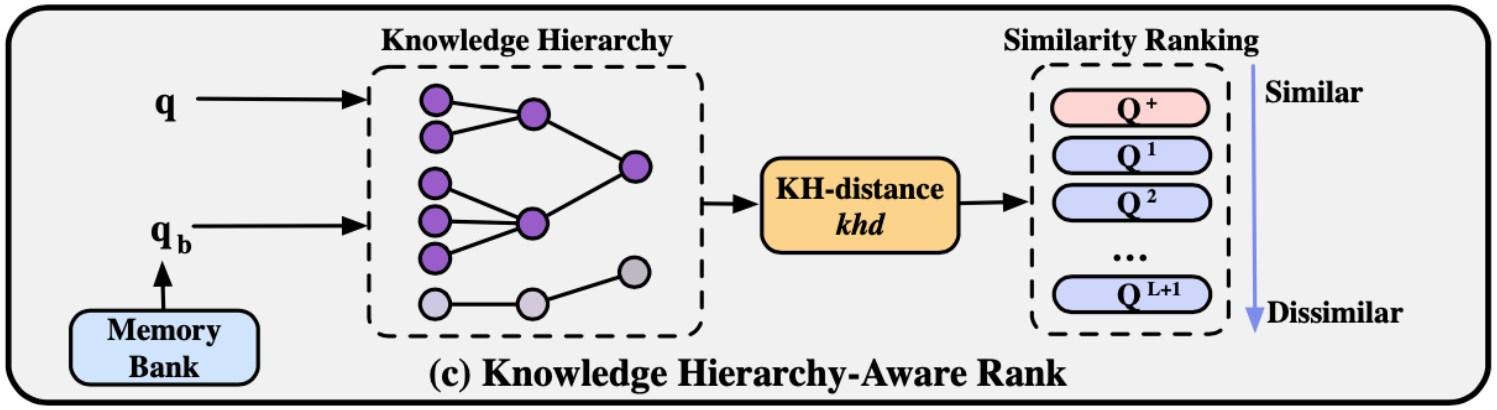
QuesCo Framework

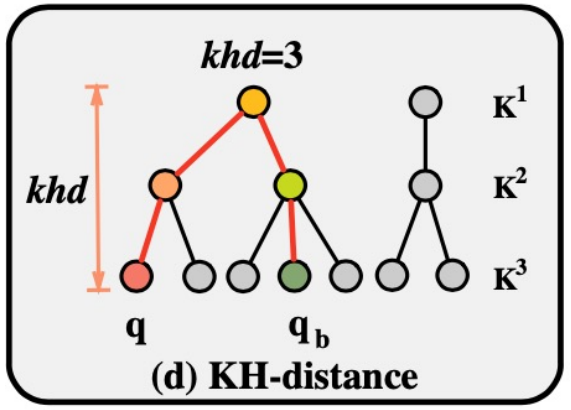
- Question Augmentation
  - Text Augmentation
    - Adopt two commonly used strategies for plain texts, i.e., random swap and random deletion
    - Not all strategies for general texts can be directly applied, such as synonym replacement
  - Formula Augmentation
    - Variable renaming
    - Variable scaling
    - Operator synonym replacement
    - Number replacement
  - Structure Augmentation
    - Clause shuffling
    - Useless clause insertion



QuesCo Framework

- Knowledge Hierarchy-Aware Rank
  - Knowledge concepts are important to understand mathematical questions
    - reflect the mathematical domain and holistic purposes
  - Knowledge concepts are not independent and contain rich information in their complex structure
    - mathematical question pairs cannot be simply divided into positives and negatives based on them
  - **Goal:** Exploit fine-grained similarities between questions based on the relationship of mathematical knowledge concepts





QuesCo Framework

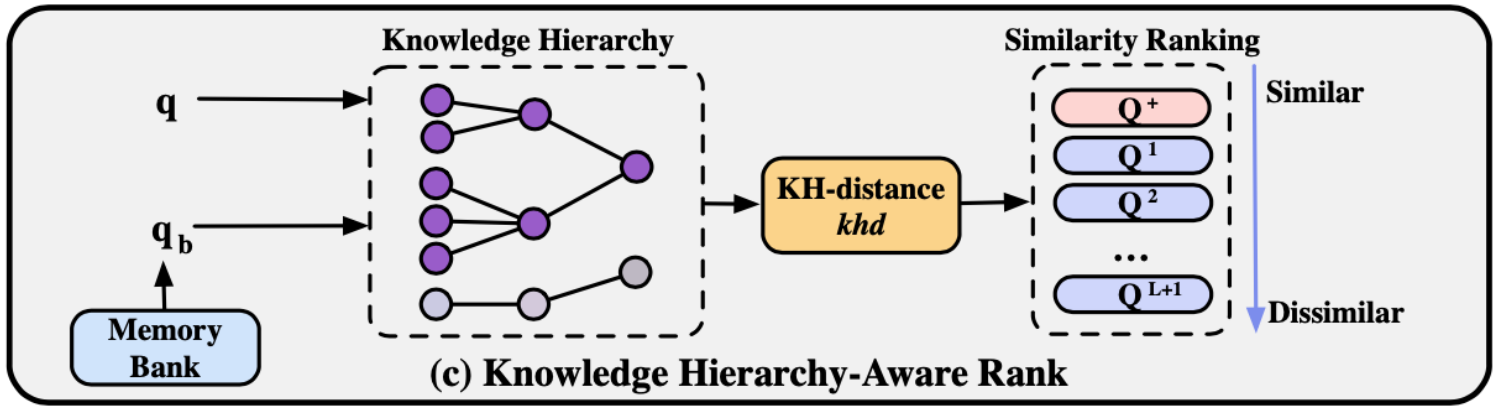
- Knowledge Hierarchy-Aware Rank
  - KH-distance

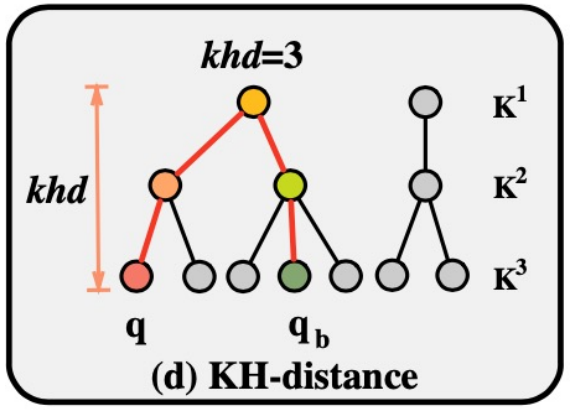
$$khd(q_i, q_j) = \begin{cases} L - u + 1 & \text{if } \exists u \in \{1, \cdots, L\}, \\ & k_i^u = k_j^u \text{ and } k_i^{u+1} \neq k_j^{u+1} \\ L + 1 & \text{if } \forall \ell \in \{1, \cdots, L\}, k_i^\ell \neq k_j^\ell \end{cases},$$

  - Assign each question q_i in memory bank into one of $L + 1$ ranks

$$Q^u = \begin{cases} \{q^+\}, & u = 0 \\ \{p | khd(q, p) = u\}, & u \in \{1, \cdots, L+1\} \end{cases}.$$

  - Similarity ranking $h(q, q^0) > h(q, q^1) > \cdots > h(q, q^{L+1}), \forall q^u \in Q^u,$





QuesCo Framework

- Pre-training
  - Ranking contrastive loss
    - Ranking Info Noise Contrastive Estimation loss (RINCE)

$$L_{rank} = \sum_0^L \ell_i$$

$$\ell_i = -\log \frac{\sum_{p \in Q^i} \exp\left(\frac{h(q,p)}{\tau_i}\right)}{\sum_{p \in \bigcup_{j \geq i} Q^j} \exp\left(\frac{h(q,p)}{\tau_j}\right)}.$$

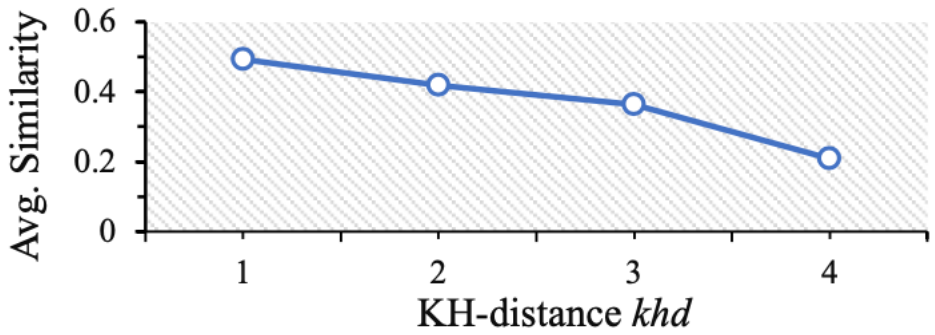
  - Gradually decreasing similarity with increasing rank of samples

Outline

Experiment

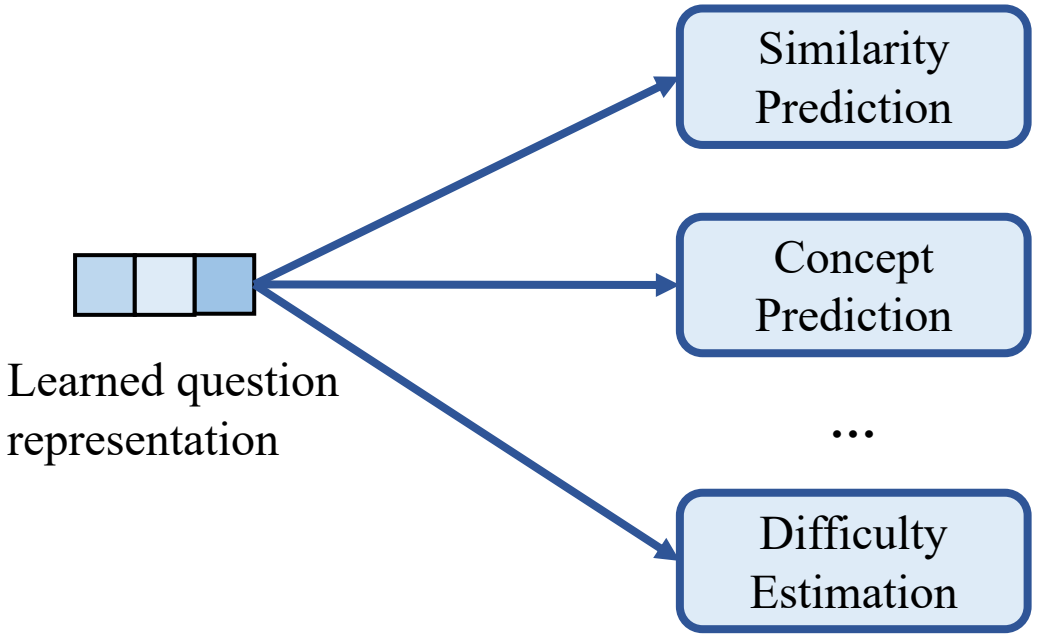
- Dataset
  - SYSTEM1
    - High school mathematical questions from an online learning system Zhixue
    - Calculate the correct rates of students as the difficulty scores of questions
  - SYSTEM2
    - High school mathematical questions from exams and textbooks, and is labeled by experts
    - Invite three experts to label similarities of question pairs

Statistics	SYSTEM1	SYSTEM2
# Questions	123,200	25,287
Avg. question length	80.08	130.91
Avg. formula length per question	48.02	84.48
# Hierarchical levels	3	3
# Knowledge in level-1	21	21
# Knowledge in level-2	81	54
# Knowledge in level-3	361	175
# Questions with difficulty label	7,056	/
# Questions with similarity label	/	6,873
Label sparsity	5.72%	27.18%



Experiment

- Evaluation Tasks
  - Similarity prediction
  - Concept prediction
  - Difficulty Estimation
- Comparison Methods
  - General pre-training method
    - BERT, DAPT-BERT
  - Contrastive learning method
    - ConSERT, SCL
  - Question pre-training method
    - QuesNet, DisenQNet



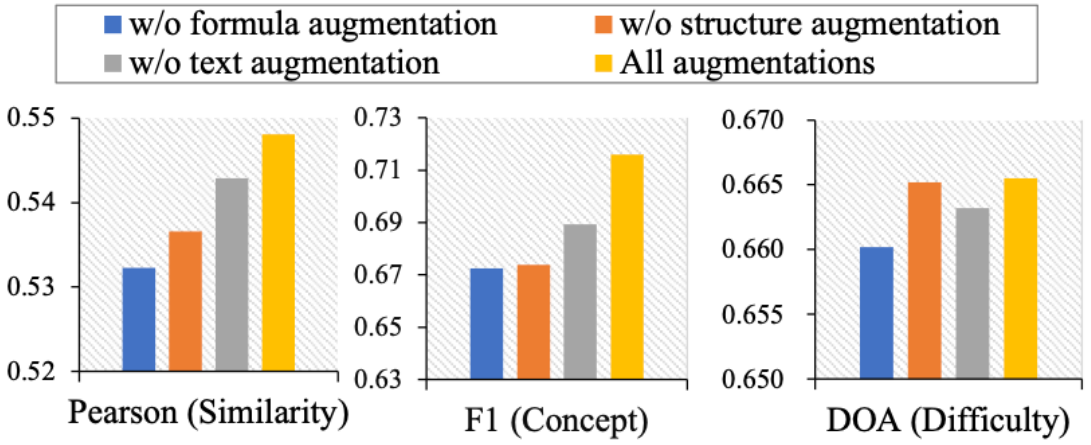
Experiment

- Observation
  - QuesCo reaches the best when predicting similarity: the modeling of the complex similarity relationship is effective
  - QuesCo performs better than knowledge-enhanced baselines: introducing the knowledge hierarchy is effective for capturing subtle differences
  - QuesCo performs well when predicting difficulty: the domain knowledge incorporated by question augmentations is efficient in capturing the difficulty embedded in diverse questions

Tasks	Similarity Prediction		Concept Prediction								Difficulty Estimation			
Datasets	SYSTEM2		SYSTEM1				SYSTEM2				SYSTEM1			
Metrics	Pearson	Spearman	level-1		level-2		level-1		level-2		MAE	RMSE	PCC	DOA
			ACC	F1	ACC	F1	ACC	F1	ACC	F1				
BERT	0.2957	0.3655	0.7309	0.5213	0.4472	0.1833	0.4822	0.2374	0.2984	0.0945	0.1987	0.2463	0.3974	0.6318
DAPT-BERT	0.4856	0.5313	0.8032	0.6288	0.5597	0.2727	0.6522	0.3855	0.4960	0.1836	0.1880	0.2313	0.5087	0.6589
ConSERT	0.5060	0.4760	0.8064	0.6655	0.5933	0.3135	0.6987	0.4952	0.5020	0.2076	0.1873	0.2308	0.5115	0.6621
SCL	0.6901	0.7101	0.8985	0.8011	0.7492	0.4683	0.8083	0.6498	0.6225	0.3071	0.1996	0.2460	0.4002	0.6340
QuesNet	0.5370	0.5549	0.7881	0.6930	0.5693	0.3604	0.7194	0.6213	0.5810	0.3118	0.1865	0.2305	0.3959	0.6539
DisenQNet	0.6922	0.6955	0.8210	0.7064	0.6404	0.4332	0.7945	0.6805	0.2431	0.1023	0.1970	0.2424	0.4293	0.6338
QuesCo	**0.7385**	**0.7245**	**0.9176**	**0.8938**	**0.7857**	**0.5550**	**0.8340**	**0.7018**	**0.6719**	**0.3756**	**0.1778**	**0.2219**	**0.5623**	**0.6765**

Experiment

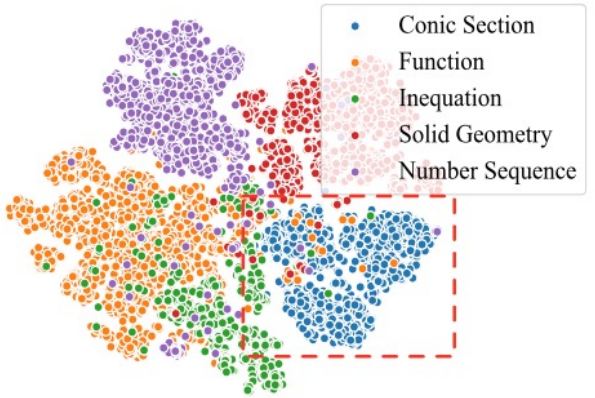
- Model Analysis
  - Ablation Study
    - Ablation on different modules
    - Ablation on different augmentation strategies
  - Observation
    - Removing any module will lead to a performance decrease on downstream tasks
    - Performance decreases when any group of augmentation strategies is removed



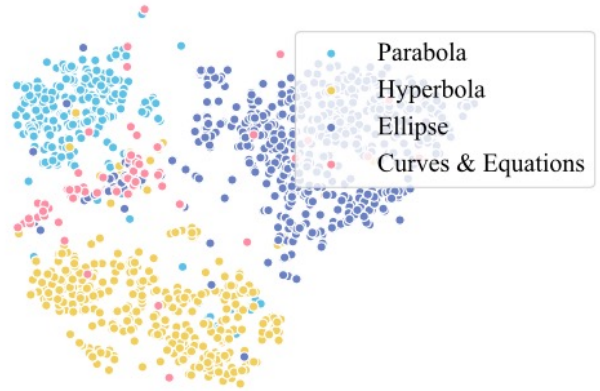
Tasks	Similarity Prediction		Concept Prediction								Difficulty Estimation			
Datasets	SYSTEM2		SYSTEM1				SYSTEM2				SYSTEM1			
Metrics	Pearson	Spearman	level-1		level-2		level-1		level-2		MAE	RMSE	PCC	DOA
			ACC	F1	ACC	F1	ACC	F1	ACC	F1				
QuesCo	**0.7385**	**0.7245**	**0.9176**	**0.8938**	**0.7857**	**0.5550**	**0.8340**	**0.7018**	**0.6719**	**0.3756**	**0.1778**	**0.2219**	**0.5623**	**0.6765**
w/o AUG	0.7028	0.7213	0.9079	0.8770	0.7305	0.4497	0.8320	0.6972	0.6443	0.3412	0.2007	0.2482	0.3797	0.6204
w/o KHAR	0.5481	0.5057	0.8202	0.7160	0.6181	0.3416	0.7332	0.5996	0.5613	0.2746	0.1810	0.2248	0.5475	0.6655

Experiment

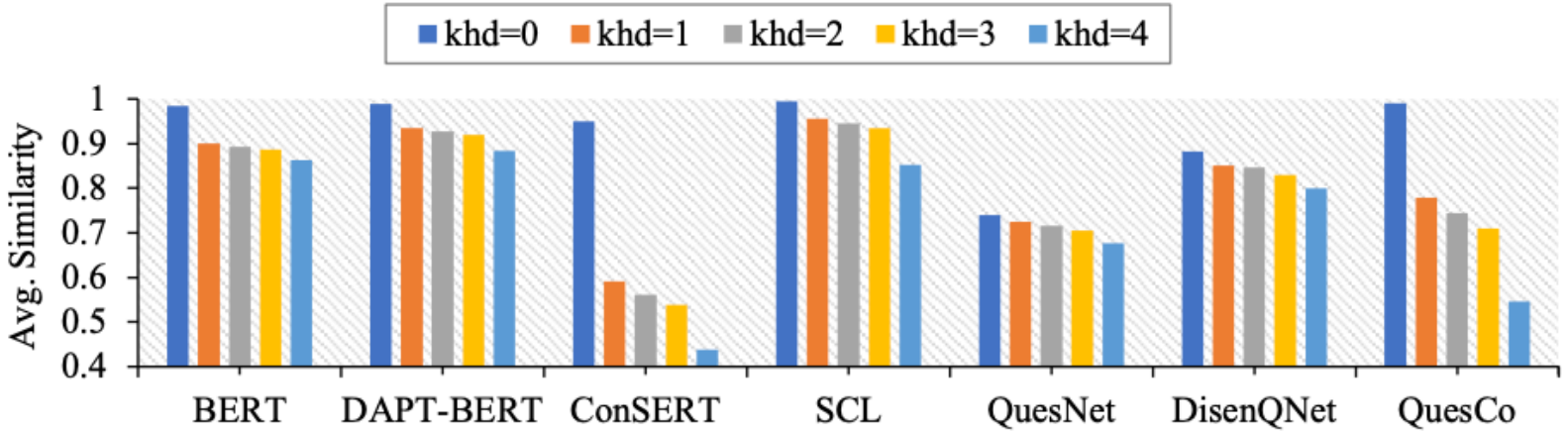
- Model Analysis
  - Visualization
    - Questions with same concepts are easy to be grouped
    - Questions with the same level-2 concepts are also grouped
  - Similarity Ranking Analysis
    - The differences of similarities obtained by QuesCo are more obvious across different ranks.



(a) level-1 knowledge concepts



(b) level-2 knowledge concepts



Experiment

- Model Analysis
  - Case Study
    - The similarity between Q_1 and the augmented question Q_2 is the highest.
    - As the KH-distance khd increases, the similarities to Q1 decrease.

Question	Knowledge Concept	*khd*	Simil-arity
Q_1: Given that $x = 1.5^{-0.2}$, $y = 1.3^{0.7}$, $z = \left(\frac{2}{3}\right)^{\frac{1}{3}}$. What is the relationship between the magnitude of x, y, z?	Exponential Function	/	/
Q_2: Given that $u = 1.5^{-0.2}$, $z = \left(\frac{2}{3}\right)^{\frac{1}{3}}$, $y = 2.3^{0.7}$. What is the relationship between the magnitude of u, y, z?		0	0.96
Q_3: If $f(x) = e^x - 1$, $g(x) = -x^2 + 4x - 3$, and $f(a) = g(b)$. What the range of value of b?		1	0.70
Q_4: What is the domain of definition of the function $f(x) = \ln(x^2 - x)$?	Logarithmic Function	2	0.64
Q_5: $f(x) = (m^2 - m - 1) \cdot x^m$ ($m \in R$) is a power function and is increasing when $x \in (0, +\infty)$. What is the value of m?	Power Function	3	0.54
Q_6: If 2 cards are randomly selected from a mixed deck (52 cards in total), what is the probability of "both are Hearts"?	Independen-ce of Events	4	0.42

Outline

Conclusion & Future work

- Conclusion
  - Study the problem of mathematical question understanding
  - Propose a novel contrastive pre-training approach, namely QuesCo, to holistically understand mathematical questions
    - Design two-level augmentations for the content and structure of mathematical questions
    - Propose a novel knowledge hierarchy-aware rank strategy to exploit the fine-grained similarity ranking between questions
  - Demonstrate the effectiveness of QuesCo with extensive experiments

- Future work
  - Generalize our work to more educational questions and exploit more educational properties
  - Design more intelligent question-based applications for personalized education

Q & A

Thanks!

ningyt@mail.ustc.edu.cn