QuesNet: A Unified Representation for Heterogeneous Test Questions

Yu Yin¹, Qi Liu¹, Zhenya Huang¹, Enhong Chen¹*, Wei Tong¹, Shijin Wang²,³, Yu Su²,³

¹Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, 
{yxonic,huangzhy}@mail.ustc.edu.cn, {cheneh,qiliuql}@ustc.edu.cn
²iFLYTEK Research, iFLYTEK CO., LTD., ³State Key Laboratory of Cognitive Intelligence, {sjwang3,yusu}@iflytek.com

Abstract
Understanding learning materials (e.g. test questions) is a crucial issue in online learning systems. Unfortunately, many supervised approaches suffer from the problem of scarce human labeled data, whereas abundant unlabeled resources are highly underutilized. While pretraining has the ability to alleviate this problem, existing pretraining methods in NLP area are infeasible to learn test question representations due to several problems. First, questions usually comprise of heterogeneous data including content text, images and side information. Second, there exists both basic linguistic information as well as domain logic and knowledge. To this end, here we propose a novel pre-training method, namely QuesNet. We first design a unified framework to aggregate question information with its heterogeneous inputs into a comprehensive vector. Then we propose a two-level hierarchical pre-training algorithm consisting of a novel holed language model objective and a domain-oriented objective, to learn better understanding of test questions. We conduct extensive experiments on large-scale real-world question data, where the experimental results clearly demonstrate the effectiveness of QuesNet for question understanding as well as its superior applicability.

Introduction

Background
How to get better question understanding is the fundamental issue promoting many question-based applications, such as difficulty estimation, knowledge mapping and score prediction.

Existing Approaches
• Rules for syntactic p:atterns or semantic encodings
• Representation networks in end-to-end frameworks

Pre-training
Take full advantage of large-scale unlabeled question corpus

Challenges
• Test questions contain coherent heterogeneous data
• We need to carefully consider the advanced logic information aside from the basic linguistic context
• The learned question representations should have great accessibility and be easy to apply to downstream tasks

QuesNet: Modeling and Pre-training

QuesNet Architecture
The overall architecture of QuesNet consists of three hierarchical layers:
1. Embedding layer: encodes heterogeneous information
2. Content layer: models linguistic relation and context
3. Sentence layer: aggregate linguistic information focusing on long-term and global complex relations

Pre-training
For solving previously mentioned challenges, we design a novel hierarchical pre-training algorithm:
1. Embedding pretraining: pre-train separately
2. Holed Language Model (HLM): like LM, but the probability of an input is conditioned by its context from both sides
3. Domain-Oriented Objective: use answers and options of a test question as a natural guidance

Experiments

Performance
• Pre-training based methods outperform original ones in general;
• QuesNet has the best performance guarantee among all the pre-training based methods.

Ablation Experiments
The result shows the importance of heterogeneous information, and also how each proposed techniques contribute to question understanding.

Datasets
The dataset we used, along with the large question corpus, are supplied by iFLYTEK Co., Ltd., from their online education system called Zhixue.

Evaluation Tasks
Knowledge mapping: a multi-label task
Difficulty estimation: estimate a numerical
Score prediction: a sophisticated domain-specific application

Comparison Methods

Abstract
Understanding learning materials (e.g. test questions) is a crucial issue in online learning systems. Unfortunately, many supervised approaches suffer from the problem of scarce human labeled data, whereas abundant unlabeled resources are highly underutilized. While pretraining has the ability to alleviate this problem, existing pretraining methods in NLP area are infeasible to learn test question representations due to several problems. First, questions usually comprise of heterogeneous data including content text, images and side information. Second, there exists both basic linguistic information as well as domain logic and knowledge. To this end, here we propose a novel pre-training method, namely QuesNet. We first design a unified framework to aggregate question information with its heterogeneous inputs into a comprehensive vector. Then we propose a two-level hierarchical pre-training algorithm consisting of a novel holed language model objective and a domain-oriented objective, to learn better understanding of test questions. We conduct extensive experiments on large-scale real-world question data, where the experimental results clearly demonstrate the effectiveness of QuesNet for question understanding as well as its superior applicability.

Introduction

Background
How to get better question understanding is the fundamental issue promoting many question-based applications, such as difficulty estimation, knowledge mapping and score prediction.

Existing Approaches
• Rules for syntactic p:atterns or semantic encodings
• Representation networks in end-to-end frameworks

Pre-training
Take full advantage of large-scale unlabeled question corpus

Challenges
• Test questions contain coherent heterogeneous data
• We need to carefully consider the advanced logic information aside from the basic linguistic context
• The learned question representations should have great accessibility and be easy to apply to downstream tasks

QuesNet: Modeling and Pre-training

QuesNet Architecture
The overall architecture of QuesNet consists of three hierarchical layers:
1. Embedding layer: encodes heterogeneous information
2. Content layer: models linguistic relation and context
3. Sentence layer: aggregate linguistic information focusing on long-term and global complex relations

Pre-training
For solving previously mentioned challenges, we design a novel hierarchical pre-training algorithm:
1. Embedding pretraining: pre-train separately
2. Holed Language Model (HLM): like LM, but the probability of an input is conditioned by its context from both sides
3. Domain-Oriented Objective: use answers and options of a test question as a natural guidance

Experiments

Performance
• Pre-training based methods outperform original ones in general;
• QuesNet has the best performance guarantee among all the pre-training based methods.

Ablation Experiments
The result shows the importance of heterogeneous information, and also how each proposed techniques contribute to question understanding.

Datasets
The dataset we used, along with the large question corpus, are supplied by iFLYTEK Co., Ltd., from their online education system called Zhixue.

Evaluation Tasks
Knowledge mapping: a multi-label task
Difficulty estimation: estimate a numerical
Score prediction: a sophisticated domain-specific application

Comparison Methods