Data-driven Item Selection and Generation in Assessments and Learning

Andrew Lan

UMass Amherst
Manning College of Information & Computer Sciences
New Work: Open-ended Knowledge Tracing

Old
Student responses are **binary-valued**, correct/incorrect. Items are characterized by a few parameters: difficulty, scale

New
Use large language models (LMs) to encode item statements and generate **knowledge-guided, open-ended** response predictions

Proof-of-concept: programming exercises dataset

Liu, Wang, Baraniuk, and Lan, EMNLP 2022
OKT can make **personalized** predictions of **structure and approach** in actual student code.
Write a function in Java that implements the following logic:
Your cell phone rings. Return true if you should answer it. Normally you answer, except in the morning you only answer if it is your mom calling. In all cases, if you are asleep, you do not answer.

```java
public boolean answerCell(boolean isMorning, boolean isMom, boolean isAsleep) {
    if (isAsleep) {
        return false;
    } else if (isMom) {
        return true;
    } else {
        return isMorning;
    }
}
```
Back to the Talk: Assessment and Learning

Assessment and learning can now happen simultaneously.
Digital Learning Platforms

Opportunity

Digital learning platforms (DLPs) produce large-scale learning data and enables personalization.
Data-driven Item Selection and Generation

- **Item selection**
  - **Bilevel optimization** based computerized adaptive testing

  Ghosh and Lan, IJCAI 2021

- **Item generation**
  - **Controlled generation** of math word problems using language models

<table>
<thead>
<tr>
<th>Equation</th>
<th>Context</th>
<th>Gen. MWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>x = num1 - num2</td>
<td>cousin game playing points scored video</td>
<td>Zach scored num1 points in the football game. Ben scored num2 points. How many more points did Zach score than Ben?</td>
</tr>
</tbody>
</table>
Computerized Adaptive Testing (CAT)

Goal

Reduce test length needed to accurately measure learner ability
How CAT Works

Select **personalized** next question for each test taker from a **question bank**
How CAT Works

Select **personalized** next question for each test taker from a **question bank**
How CAT Works

- Estimate test taker **ability** using item response theory (IRT)
- Select next item that provides most **information** on ability
Simple yet powerful

\[ p(Y_{i,j} = 1) = \sigma(\theta_i - b_j), \]

The 1PL IRT model

- \( Y_{i,j} \in \{0, 1\} \): binary-valued question response correctness
- \( \sigma(\cdot) : \mathbb{R} \to [0, 1] \): the sigmoid function

There are many variants; we use 1PL as a running example
How CAT Works

At the end of the test, score according to the ability estimate.

- Fixed-length and varied-length CAT.
Nature of CAT

- If the test contains every question in the question bank, that score would be highly accurate.
- But that is not practical.
- Ability is a **proxy** for score on that long test.
Some problems with CAT

- IRT is not the most flexible and predictive model
- The informativeness metric is static

Opportunity

Models and item selection algorithms that can truly exploit large-scale learner response data
Goal: learn a **data-driven** question selection algorithm and support **flexible response models**
We solve the following bilevel optimization problem:

$$\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Gamma_i} \ell \left( Y_{i,j}, g(j; \theta_{i}^*) \right) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\theta_{i}^*, \Gamma_i) \\
\text{s.t.} & \quad \theta_{i}^* = \arg \min_{\theta_i} \sum_{t=1}^{n} \ell \left( Y_{i,j_i(t)}, g(j_{i}^{(t)}; \theta_i) \right) + \mathcal{R}(\gamma, \theta_i) := \mathcal{L}'(\theta_i) \\
\text{where} & \quad j_{i}^{(t)} \sim \Pi(Y_{i,j_{i}^{(1)}}, \ldots, Y_{i,j_{i}^{(t-1)}}; \phi) \in \Omega_{i}^{(t)}
\end{align*}$$

- (1): outer optimization problem, learn **global response model** parameters $\gamma$ and **item selection algorithm** parameters $\phi$
- (2): inner optimization problem, learn **local response model** parameters for each test taker $\theta_{i}^*$
- (3): item selection algorithm
Meta-learning setup that splits data into training and meta sets
BOBCAT Details

- **Agnostic** to the response model: IRT/neural network
- **Global-local** parameter split in the response model - flexible, one example:
  - Global: difficulty/weights&biases
  - Local: ability/input vector
- **Gradient** calculation for global variables $\gamma, \phi$
  - Unbiased using REINFORCE
  - Biased using influence function scores (works better)
- Computational efficiency: **more** efficient than CAT
  - The item selection algorithm is a neural network with raw responses as input
  - Only need a forward pass
  - No need to update the ability estimate after each response
Experiments: Data and Metrics

Data: five large-scale, real-world learner response datasets
Metrics: accuracy and AUC on test set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>EdNet</th>
<th>Junyi</th>
<th>Eedi-1</th>
<th>Eedi-2</th>
<th>ASSISTments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>312K</td>
<td>52K</td>
<td>119K</td>
<td>5K</td>
<td>2.3K</td>
</tr>
<tr>
<td>Questions</td>
<td>13K</td>
<td>25.8K</td>
<td>27.6K</td>
<td>1K</td>
<td>26.7K</td>
</tr>
<tr>
<td>Interactions</td>
<td>76M</td>
<td>13M</td>
<td>15M</td>
<td>1.4M</td>
<td>325K</td>
</tr>
</tbody>
</table>
Experiments: Methods (selected)

- **IRT-active**: IRT as response model without BOBCAT, uncertainty sampling as item selection algorithm
- **BiIRRT-active**: IRT as response model in BOBCAT, uncertainty sampling as item selection algorithm
- **BiNN-active**: neural network as response model in BOBCAT, uncertainty item as question selection algorithm
- **BiNN-unbiased**: neural network as response model in BOBCAT, learned item selection algorithm with unbiased gradient estimate
- **BiNN-approx**: neural network as response model in BOBCAT, learned item selection algorithm with biased, approximate gradient estimate
## Experiments: Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>IRT-Active</th>
<th>BiIRT-Active</th>
<th>BiIRT-Unbiased</th>
<th>BiIRT-Approx</th>
<th>BiNN-Approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>EdNet</td>
<td>1</td>
<td>70.08</td>
<td>70.92</td>
<td>71.12</td>
<td><strong>71.22</strong></td>
<td><strong>71.22</strong></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>70.63</td>
<td>71.16</td>
<td>71.3</td>
<td>71.72</td>
<td>71.82</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>71.03</td>
<td>71.37</td>
<td>71.45</td>
<td>71.95</td>
<td><strong>72.17</strong></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>71.62</td>
<td>71.75</td>
<td>71.79</td>
<td>72.33</td>
<td>72.55</td>
</tr>
<tr>
<td>Junyi</td>
<td>1</td>
<td>74.52</td>
<td>74.93</td>
<td>74.97</td>
<td><strong>75.11</strong></td>
<td>75.1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>75.19</td>
<td>75.48</td>
<td>75.53</td>
<td>75.76</td>
<td><strong>75.83</strong></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>75.64</td>
<td>75.79</td>
<td>75.75</td>
<td>76.11</td>
<td><strong>76.19</strong></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>76.27</td>
<td>76.28</td>
<td>76.19</td>
<td>76.49</td>
<td><strong>76.62</strong></td>
</tr>
<tr>
<td>Eedi-1</td>
<td>1</td>
<td>66.92</td>
<td>68.22</td>
<td>68.61</td>
<td><strong>68.82</strong></td>
<td>68.78</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>68.79</td>
<td>69.45</td>
<td>69.81</td>
<td>70.3</td>
<td><strong>70.45</strong></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>70.15</td>
<td>70.28</td>
<td>70.47</td>
<td>70.93</td>
<td><strong>71.37</strong></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>71.72</td>
<td>71.45</td>
<td>71.57</td>
<td>72.0</td>
<td><strong>72.33</strong></td>
</tr>
<tr>
<td>Eedi-2</td>
<td>1</td>
<td>63.75</td>
<td>64.83</td>
<td>65.22</td>
<td>65.3</td>
<td><strong>65.65</strong></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>65.25</td>
<td>66.42</td>
<td>67.09</td>
<td>67.23</td>
<td><strong>67.79</strong></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>66.41</td>
<td>67.35</td>
<td>67.91</td>
<td>68.23</td>
<td><strong>68.82</strong></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>68.04</td>
<td>68.99</td>
<td>68.84</td>
<td>69.47</td>
<td><strong>70.04</strong></td>
</tr>
<tr>
<td>ASSIST3</td>
<td>1</td>
<td>66.19</td>
<td>68.69</td>
<td>69.03</td>
<td><strong>69.17</strong></td>
<td>68.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>68.75</td>
<td>69.54</td>
<td>69.78</td>
<td><strong>70.21</strong></td>
<td>68.73</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>71.04</td>
<td>70.66</td>
<td><strong>71.17</strong></td>
<td>71.14</td>
<td>69.03</td>
</tr>
</tbody>
</table>

- Data-driven item selection algorithms are better than informativeness-driven ones
- Biased approximate gradient works much better than unbiased gradient
- Neural network works better than IRT on larger datasets
Even if we still score test takers using IRT ability estimates, the learned item selection algorithm is much more effective than typical informativeness metrics.

It gets better with more training data.
Test Security Problems

<table>
<thead>
<tr>
<th>Method</th>
<th>Exposure (median)</th>
<th>Exposure (&gt;20%)</th>
<th>Overlap (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT-Active</td>
<td>0.51%</td>
<td>0.25%</td>
<td>6.03%</td>
</tr>
<tr>
<td>BiNN-Approx</td>
<td>0%</td>
<td>1.54%</td>
<td>28.64%</td>
</tr>
</tbody>
</table>

- Learned question selection algorithms lead to higher question exposure rates and much higher test overlap rates.
- Forthcoming work adds corresponding constraints into our optimization objective to address this problem.
BOBCAT improves CAT by learning item selection algorithms from data: the more data, the better the algorithm.

- Supports any response model
Data-driven Item Selection and Generation

- **Item selection**
  - **Bilevel optimization** based computerized adaptive testing

- **Item generation**
  - **Controlled generation** of math word problems using language models

Wang, Baraniuk, and Lan, EMNLP 2021
Personalizing Items

- Even if we know how to select items perfectly, they still come from a **finite item pool**

**Standardization and UNDERSTANDarization in Educational Assessment**

Stephen G. Sireci, *University of Massachusetts Amherst*

**Problem**

- Items may not match learner interests or be **culturally relevant**

- Manually generating personalized items is helpful but not a scalable approach
**Math Word Problems (MWPs)**

**MWP:** Joan found 70 seashells on the beach. She gave Sam some of her seashells. She has 27 seashells. How many seashells did she give to Sam?

**Equation:** \( x = (70 - 27) \)

- Textual statement, underlying equation, relatively simple
- Can use **language models** (LMs), e.g., GPT-3, PaLM, to automatically generate
- But these black-box models are not controllable

**Goal**

Controllable MWP generation by specifying **equation** & **context**
Controllable MWP Generation

\[ x = (6 - 2) \]

Joanna, balloons

Joanna has six balloons. Two popped. How many does she have left?
Controllable MWP Generation

\[ x = (6 \times 2) + 3 \]

Joanna, balloons

Joanna bought two sets of six balloons. She then bought three more. How many balloons does she have now?
Mike has six **apples** in each of the two **baskets** and three **apples** on the table. How many **apples** does he have in total?
Key Technique: Equation Consistency Control

\[ p_\phi(\text{eq} \mid \text{MWP}) = \prod_{t=1}^{T} p_\phi(e_t \mid \text{MWP}, e_{<t}) \]

\[ \mathcal{L}_{\text{eq}} = -\log p_\phi(\text{eq} \mid \text{MWP}) \]

"Gumbel-softmax to enable gradient propagation"

Context: Joanna, balloons

Equation: \[ x = (6 - 2) \]

MWP generator

Generated MWP

Equation generator

Predicted equation
Key Technique: Context Keyword Selection

\[
\mathcal{L}_c = \text{KL}(q_\psi \parallel p_c)
\]

\[
q_\psi(w_i) = \sigma(a_i) \mathbf{1}_{\{w_i \in \text{MWP}\}}
\]

Only select words from a given MWP

\[
\mathcal{L}_{\text{eq}} = -\log p_\phi(\text{eq} \mid \text{MWP})
\]
Quantitative Results

<table>
<thead>
<tr>
<th></th>
<th>Arithmetic</th>
<th>MAWPS</th>
<th>Math23K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-4</td>
<td>METEOR</td>
<td>ROUGE-L</td>
</tr>
<tr>
<td>seq2seq-rnn</td>
<td>0.075</td>
<td>0.152</td>
<td>0.311</td>
</tr>
<tr>
<td>+ GloVe</td>
<td>0.351</td>
<td>0.310</td>
<td>0.555</td>
</tr>
<tr>
<td>seq2seq-tf</td>
<td>0.339</td>
<td>0.298</td>
<td>0.524</td>
</tr>
<tr>
<td>GPT</td>
<td>0.237</td>
<td>0.248</td>
<td>0.455</td>
</tr>
<tr>
<td>GPT-pre</td>
<td>0.316</td>
<td>0.322</td>
<td>0.554</td>
</tr>
<tr>
<td>ours</td>
<td>0.338</td>
<td>0.322</td>
<td>0.567</td>
</tr>
</tbody>
</table>

Three MWP datasets

Baselines: sequence-to-sequence, fine-tuning GPT

Metrics: BLEU-4, METEOR, ROUGE-L (language quality), ACC-eq (equation consistency)

Baselines tend to not really “learn” how to generate MWPs
Qualitative Results: Fixed Context, Varying Equation

Context: candies

**Equation #1:** \( x = num1 + num2 \)

- seq2seq-tf: Ethan has num1 presents. Alissa has num2 more than Ethan. How many presents does Alissa have? *(in training data)*
- GPT-pre: There are num1 scissors in the drawer. Keith placed num2 scissors in the drawer. How many scissors are now there in total? *(irrelevant to context)*
- ours: Mildred collects num1 candies. Mildred’s father gives Mildred num2 more. How many candies does Mildred have? *(✓)*

**Equation #2:** \( x = num1 - num2 \)

- seq2seq-tf: Mildred weighs num1 pounds. Carol weighs num2 pounds. How much heavier is Mildred than Carol? *(in training data)*
- GPT-pre: Joan has num1 blue balloons but lost num2 of them. How many blue balloons does Joan have now? *(irrelevant to context)*
- ours: There are num1 candies in the jar. num2 are eaten by a hippopotamus. How many candies are in the jar? *(✓)*

**Equation #3:** \( x = num1 \times num2 \)

- seq2seq-tf: Each banana costs $ num1. How much do num2 bananas cost? *(in training data)*
- GPT-pre: Joan has saved num1 quarters from washing cars. How many cents does Joan have? *(inconsistent with equation)*
- ours: Each child has num1 candies. If there are num2 children, how many candies are there in all? *(✓)*

**Equation #4:** \( x = num1 / num2 \)

- seq2seq-tf: There are num1 bananas in Diane’s banana collection. If the bananas are organized into num2 groups, how big is each group? *(in training data)*
- GPT-pre: Joan has num1 blue marbles. Sandy has num2 times more blue marbles than Melanie. How many blue marbles does Joan have? *(inconsistent with equation)*
- ours: There are num1 candies in the candy collection. If the candies are organized into num2 groups, how big is each group? *(✓)*
### Qualitative Results: Fixed Equation, Varying Context

**Equation:** \( x = \text{num1} + \text{num2} + \text{num3} \)

<table>
<thead>
<tr>
<th>Context #1: violin piano acoustic guitar</th>
<th>Context #2: beets eggplant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>seq2seq-tf:</strong> sara grew num1 onions, sally grew num2 onions, and fred grew num3 onions. how many onions did they grow in all? (in training data)</td>
<td><strong>seq2seq-tf:</strong> sara grew num1 onions, sally grew num2 onions, and fred grew num3 onions. how many onions did they grow in all? (in training data)</td>
</tr>
<tr>
<td><strong>GPT-pre:</strong> There are num1 dogwood trees currently in the park. Park workers will plant num2 dogwood trees today and num3 dogwood trees tomorrow. How many dogwood trees will the park have when the workers are finished? (irrelevant to context)</td>
<td><strong>GPT-pre:</strong> There are num1 orchid bushes currently in the park. Park workers will plant num2 orchid bushes today and num3 orchid bushes tomorrow. How many orchid bushes will the park have when the workers are finished? (irrelevant to context)</td>
</tr>
<tr>
<td><strong>ours:</strong> Mike joined his school’s band. He bought a clarinet for $\text{num1}$, a music stand for $\text{num2}$, and a song book for $\text{num3}$. How much did Mike spend at the music store? (✓)</td>
<td><strong>ours:</strong> Sara grew num1 beets, Sally grew num2 beets, and Fred grew num3 beets. How many beets did they grow in total? (✓)</td>
</tr>
</tbody>
</table>
We provide a method for the context/equation-controllable MWP generation method based on LMs.

LMs have the ability to adapt to many topics due to their intrinsic knowledge.

However, LMs are not very good at mathematical reasoning.

The safety of the generated text also needs to be monitored.
The Future

- Contrary to popular belief, big data in education is still in its early stages.
- Need to develop ways for human and artificial intelligence to work together.