Generalizing Expert Misconception Diagnoses Through Common Wrong Answer Embedding

John Kolb, Scott Farrar, (presenter) Zachary A. Pardos

UC Berkeley
Our Goal

• Automatically generate the student misconceptions underlying a commonly occurring wrong answer
Our Goal

• Automatically generate the student misconceptions underlying a commonly occurring wrong answer

• **Misconception**: An incorrect understanding of a topic leading to false conclusions
  - Rich treatment in literature
  - Can be difficult to work backwards from answer to its origins in the student's misunderstanding
Potential Impact on teaching

Teacher dashboard depicting which misconception their students are close to

Observations:

(1) There are a lot of student problem solving data

(2) There are not a lot of semantic diagnostic data
This training process, using SGD, is run on a corpus of 1b words to learn vector representations of each word in the vocabulary.

“All happy families are alike; each unhappy family is unhappy in its own way.”

Mikolov, T., & Dean, J. (2013)
Related work

1. Problem2vec used to predict skill of item (Dadu & Pardos, 2017)
2. Course2vec used to find university course similarity (Pardos, Jiang, & Fan, 2019)
3. Initial pilot study of misconceptions using w2v (Pardos et al., 2018)
4. Tree or graph-based representations of student problem-solving “maps” (Eagle & Barnes, 2014; Muehling, 2017)
5. Inferring misconceptions from open-text (Michalenko, Lan, & Baraniuk, 2017)
Data Set (Khan Academy)

• Each student answer log contained:
  o *Exercise*: Overall topic/skill
  o *Problem Type*: Generic question template
  o *Seed*: Unique template instantiation
  o *Timestamp*
Data Set (Khan Academy)

• Each student answer log contained:
  o **Exercise**: Overall topic/skill
  o **Problem Type**: Generic question template
  o **Seed**: Unique template instantiation
  o **Timestamp**

selected for likelihood of misconceptions:

"Surface areas"
"Slope from an equation in slope-intercept form"
"Area of quadrilaterals and polygons"
"Adding and subtracting fractions"
## Data Set (Khan Academy)

<table>
<thead>
<tr>
<th>Problem Types</th>
<th>Surface Areas</th>
<th>Slope-Intercept</th>
<th>Area of Quads.</th>
<th>Add/Sub Fractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeds</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique Users</td>
<td>105,659</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique Incorrect Answers</td>
<td>55,126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Incorrect Answers</td>
<td>619,045</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Data Set (Khan Academy)

<table>
<thead>
<tr>
<th>Problem Types</th>
<th>Surface Areas</th>
<th>Slope-Intercept</th>
<th>Area of Quads.</th>
<th>Add/Sub Fractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeds</td>
<td>38</td>
<td>20</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Unique Users</td>
<td>105,659</td>
<td>33,603</td>
<td>58,239</td>
<td>179,263</td>
</tr>
<tr>
<td>Unique Incorrect Answers</td>
<td>55,126</td>
<td>6,912</td>
<td>17,998</td>
<td>46,516</td>
</tr>
<tr>
<td>Total Incorrect Answers</td>
<td>619,045</td>
<td>112,390</td>
<td>298,356</td>
<td>873,916</td>
</tr>
</tbody>
</table>
Approach Summary

Incorrect Student Answers
(Khan Academy)
Approach Summary

Incorrect Student Answers (Khan Academy)

Tokenization

Student Answer Sequences

["a_1", c_3", ..."]
Approach Summary

Tokenization

Student Answer Sequences

["a_1", c_3", …]

Skip-Gram Model

Token Embeddings

"a_1" = [0.24, 1.41, …]

Incorrect Student Answers (Khan Academy)
Approach Summary

Incorrect Student Answers (Khan Academy)

Tokenization

Student Answer Sequences

["a_1", c_3", ...]

Skip-Gram Model

Token Embeddings

"a_1" = [0.24, 1.41, ...]
Approach Summary

Incorrect Student Answers (Khan Academy)

Tokenization

Student Answer Sequences

["a_1", c_3", ...]

Skip-Gram Model

Token Embeddings

"a_1" = [0.24, 1.41, ...]

Expert Misconception Analysis

"Student forgot a factor of 2 in formula"
Approach Summary

Tokenization

Incorrect Student Answers (Khan Academy)

Student Answer Sequences

"a_1", c_3", …"

Skip-Gram Model

Token Embeddings

"a_1" = [0.24, 1.41, …]

Expert Misconception Analysis

Misconception Diagnoses

"Student forgot a factor of 2 in formula"

Postprocessing

Bag of Words Representations

[0, 1, 0, 0, 1,…]
Approach Summary

Tokenization: Tokenization sequences

Incorrect Student Answers (Khan Academy):

Expert Misconception Analysis: "Student forgot a factor of 2 in formula"

Skip-Gram Model: Token Embeddings

Misconception Diagnoses:

Postprocessing: Bag of Words Representations

Sequences: ["a_1", c_3", …]
Approach Summary

Tokenization

Student Answer Sequences
["a_1", c_3", ...]

Skip-Gram Model

Token Embeddings
"a_1" = [0.24, 1.41, ...]

Multinomial Logistic Regression

Predicted Misconception Tags
[0.01, 0.035, 0.11, ...]

Postprocessing

Bag of Words Representations
[0, 1, 0, 0, 1, ...]

Incorrect Student Answers (Khan Academy)

Expert Misconception Analysis

"Student forgot a factor of 2 in formula"

Misconception Diagnoses

"a_1" = [0.24, 1.41, ...]
Computing Embeddings – Skip-Gram Model

• Why skip-gram?
  o Surfaces insights about words based on the *context* in which they appear (surrounding words in a sentence)
Computing Embeddings – Skip-Gram Model

• Why skip-gram?
  o Surfaces insights about words based on the context in which they appear (surrounding words in a sentence)

The quick brown fox jumps over the lazy dog.
Computing Embeddings – Skip-Gram Model

• Why skip-gram?
  o Surfaces insights about words based on the context in which they appear (surrounding words in a sentence)
  
  The quick brown fox jumps over the lazy dog.

  o We seek insights about incorrect answers based on a similar context (chronologically preceding and following answers)
Computing Embeddings – Skip-Gram Model

- Why skip-gram?
  - Surfaces insights about words based on the context in which they appear (surrounding words in a sentence)
  - We seek insights about incorrect answers based on a similar context (chronologically preceding and following answers)
Computing Embeddings – Skip-Gram Model

• Why skip-gram?
  - Surfaces insights about words based on the context in which they appear (surrounding words in a sentence)
  - We seek insights about incorrect answers based on a similar context (chronologically preceding and following answers)

The quick brown fox jumps over the lazy dog.

Goal: Learn a semantic space of answer vector embeddings
Computing Embeddings – Skip-Gram Model

Time

Student response sequence
Computing Embeddings – Skip-Gram Model

Time

<table>
<thead>
<tr>
<th>Seed</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>x01b</td>
<td>&quot;55&quot;</td>
</tr>
<tr>
<td>x01b</td>
<td>&quot;70&quot;</td>
</tr>
<tr>
<td>x03e</td>
<td>&quot;40&quot;</td>
</tr>
</tbody>
</table>

Student response sequence
Computing Embeddings – Skip-Gram Model

Seed: x01b
Response: “55”

Seed: x01b
Response: “7θ”

Seed: x03e
Response: “4θ”

Student response sequence:
[ "x01b_2", "x01b_6", "x0e3_c", ... ]

Skip-Gram Token Sequence
Computing Embeddings – Skip-Gram Model

Time

[ "x01b_2", "x01b_6", "x0e3_c", ... ]

Student response sequence

Seed | x01b
Response | “55”

Seed | x01b
Response | “70”

Seed | x03e
Response | “40”

Question's Unique Seed | frequency rank of incorrect answer | “correct”

Skip-Gram Token Sequence
Computing Embeddings – Skip-Gram Model

Time

Student response sequence

<table>
<thead>
<tr>
<th>Seed</th>
<th>x01b</th>
<th>Seed</th>
<th>x01b</th>
<th>Seed</th>
<th>x03e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>“55”</td>
<td>Response</td>
<td>“70”</td>
<td>Response</td>
<td>“40”</td>
</tr>
</tbody>
</table>

[ "x01b_2", "x01b_6", "x0e3_c", ... ]

Skip-Gram Token Sequence
Collecting Expert Misconception Diagnoses

• Designed and ran a survey on the Qualtrics platform
  o Participants recruited and compensated on our behalf

• Subject population:
  o Mathematics educators teaching grades 5-12 or undergraduates
  o Minimum 2 years prior experience
Collection Design

Instructions

• "Respond with a general label-phrase that describes the most likely error or misconception related to the incorrect answer."

• "You may duplicate labels and phrases as you see appropriate."

• "You will see three math questions and five incorrect student answers to label for each question."
Postprocessing of Diagnoses

• Filtered for highest quality diagnoses
  o Included work only from experts with average label character count > 20
• NLP cleaning: remove punctuation, stemming, stopword removal
  o Excluded trivially predictable words: student, tried, used, etc.
• Convert diagnosis to vector representation: Bag of Words
• Ended with 19 experts providing 570 unique diagnoses covering 14 of the 15 problem types and 64 of the 89 seeds
Approach Summary

Tokenization → Student Answer Sequences → Skip-Gram Model

Token Embeddings → Multinomial Logistic Regression

Predicted Misconception Tags → [0.01, 0.035, 0.11, ...]

Bag of Words Representations → [0, 1, 0, 0, 1, ...]

Postprocessing

Incorrect Student Answers (Khan Academy) → Expert Misconception Analysis

Misconception Diagnoses → "Student forgot a factor of 2 in formula"
Predicting/Interpolating diagnoses

• Training input: Student answer embedding ($n$-vectors) paired with an expert diagnosis ($m$-vector, bag of words) of that answer
  - $n =$ Dimensionality of embedding space, skip-gram hyperparameter
  - $m =$ Size of expert diagnosis vocabulary

• Prediction Task: Given an answer embedding, generate its misconception diagnosis
  - $m$-vector: probability distribution over diagnosis vocabulary
Evaluation: Leave-One-Out CV

<table>
<thead>
<tr>
<th></th>
<th>Evaluator</th>
<th>Problem Type</th>
<th>Seed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Folds</strong></td>
<td>19</td>
<td>14</td>
<td>64</td>
</tr>
<tr>
<td><strong>Avg. Training Data Points</strong></td>
<td>302</td>
<td>296</td>
<td>314</td>
</tr>
<tr>
<td><strong>Avg. Test Data Points</strong></td>
<td>17</td>
<td>24</td>
<td>5</td>
</tr>
</tbody>
</table>
Metric: Recall @ N

• Given:
  o An original expert diagnosis consisting of $N$ terms
  o Predicted probability distribution over the diagnosis vocabulary

• Select $N$ largest elements from probability distribution and their corresponding vocab terms

• How many of these selected terms appear in the original diagnosis?

• Formally: $R = \frac{|\hat{T}_N \cap T|}{|T|}$
Naïve Baselines for Comparison

1. **Random**: Compute a random sample of $N$ terms from the vocabulary.

2. **Frequency**: Predict the $N$ terms that appear most frequently in the diagnoses from the training set.
Results

- Our approach outperforms frequency baseline by $\sim 2x$
- Roughly 18%-27% of original terms recovered
- Accuracy rivals that of traditional word-to-word machine translation (Mikolov et al., 2013), which produced 25% accuracy translating between English and Vietnamese.
Limitations

- Open-text diagnoses had high variation
  > should limit taggers to a shared taxonomy
- W2v embedding was not cross-exercise
  > use a datasets with high exercise coverage per student
- Single word misconception “hint” may not be enough
Thank You!