Mining Misconceptions in Mathematics

Using Natural Language Processing

NLP team

BME

October 30, 2024

Table of Contents

The problem

A math teacher ask their students the following question:
What do you need to do to eliminate the qs in each of these pairs of simultaneous equations?

Pair 1	Pair 2
$2p - 4q = 6$	$2p - 4q = 6$
$7p + 4q = 9$	$7p - 4q = 9$

a) Add Pair 1 Subtract Pair 2,

b) Add Pair 1 Add Pair 2,

c) Subtract Pair 1 Add Pair 2,

d) Subtract Pair 1 Subtract Pair 2

The clear answer is a). However, the teacher is interested in knowing why some of the students decided to choose any of the wrong answers. What kind of misconceptions do the students have that made them choose wrongly?

Pair 1	Pair 2
$2p - 4q = 6$	$2p - 4q = 6$
$7p + 4q = 9$	$7p - 4q = 9$

For example, if the student chose d) Subtract Pair 1 Subtract Pair 2, we could say that the student:

Believes that when eliminating a variable, regardless of the signs of the terms with matching coefficients, we subtract the equations.

The objective

We would like to create a model that, given the text of a question and a wrong answer, it can predict the kind of misconception that may have led the student to make such a choice.

Data

We obtained a database from Kaggle (`https://www.kaggle.com/competitions/eedi-mining-misconceptions-in-mathematics/overview`.
This database contains a total of 1869 multiple choice questions, together with the following information:

- Numerical id for the question
- Type of exercise
- Subject of the exercise
- Text of the question in LaTeXformat.
- Text of the four answers in LaTeXformat.
- Correct answer
- Misconception ids for the three wrong answers*.

There are a total of 2586 misconceptions that also come in the form of LaTeXtext, together with their Ids.

Embeddings

- The most important information from our data is the text from the questions, answers and misconceptions. So we need a way to store these strings numerically (Embeddings).
- We use FAISS (Facebook AI Similarity Search), a library that allows developers to quickly search for embeddings of multimedia documents that are similar to each other.
- Basically, each string is represented by a vector, with the following property regarding the cosine similarity: $\cos(\theta_{u,v}) := \frac{u \cdot v}{\|u\|\|v\|}$
  - if $\cos(\theta_{u,v}) << 0$, then the strings corresponding to the vectors are *opposite*.
  - if $\cos(\theta_{u,v}) \approx 0$, then the strings corresponding to the vectors are *decorrelated*.
  - if $\cos(\theta_{u,v}) >> 0$, then the strings corresponding to the vectors are *similar*.

The Baseline:

- We divided our database into training and testing data.
- For an initial approach, we have considered using a Random Forest. After testing, we will move away from this choice in the future, but this is what we have now.
- For the moment, we have only used the embeddings for the questions and answers for the training data, and the tags for the corresponding Misconception Id's as labels.
- With our initial approach, we have learned that Random Forest don't handle many categories well.

Current Result

- We want our model to receive a question and a wrong answer to it. This question is external: it won't be part of the training data.
- We use the properties of the FAISS embeddings to retrieve questions **within our testing data** that are similar to the external one. We know that this is not the best decision, and we will find a different way to do it.
- We pass the predictions of the testing questions through the classifier to get the corresponding misconceptions.

```
Question: What is 8 + 12 div (2+2)? Answer: ( 12 )
 Possible misconceptions:
Retrieved Misconception 1: Misconception: Believes 0 is not a real solution to an equation
Retrieved Misconception 2: Misconception: Believes 0 is not a real solution to an equation
Retrieved Misconception 3: Misconception: Believes a number raised to a power will always give a positive answer
Retrieved Misconception 1: Misconception: Thinks the interior angles of any polygon add up to 360
Retrieved Misconception 2: Misconception: Does not know how to calculate the sum of interior angles
Retrieved Misconception 3: Misconception: Believes that you are unable to calculate the sum of the interior angles of an irregular polygon given the number of sides
Retrieved Misconception 1: Misconception: Does not understand the value of zeros as placeholders
Retrieved Misconception 2: Misconception: Rounds down instead of up
Retrieved Misconception 3: Misconception: Rounds to the wrong degree of accuracy (rounds too much)
Retrieved Misconception 1: Misconception: Forgets the denominator in probability
Retrieved Misconception 2: Misconception: Forgets the denominator in probability
Retrieved Misconception 3: Misconception: Thinks that probabilities of an event occurring or not occurring must be equal
Retrieved Misconception 1: Misconception: Does not understand that a probability of 0 represents something that is impossible
Retrieved Misconception 2: Misconception: Does not understand that a probability of 0 represents something that is impossible
Retrieved Misconception 3: Misconception: Does not understand that a probability of 0 represents something that is impossible
```

What to do for improvement:

- We have not **yet** used all the information from our data yet: Type of exercise, subject of the exercise and not even the embeddings for the misconceptions.
- We haven't performed any feature engineering.
- We need to pick a better classifier.
- We need to decide what to do with the missing data in the database (Missing misconceptions)*.
- Perform Singular Value Analysis of our data.
- We need to figure out how to update our current embeddings with *external questions* instead of directly searching for similar questions in our database (defeats the purpose of the classifier).