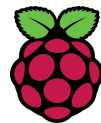
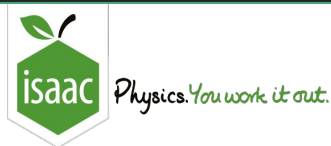


Automated marking of free-text questions in STEM

Meurig Thomas, Alastair Beresford



Department
for Education

Free-text questions assess the content of short, natural-language sentences

Answer Now

Identify the force or forces acting on the ball after it emerges from the track at R.

It's weight and the normal contact force from the table





Check my answer

Free-text questions influence the learner's response less than alternative question types

- Assesses recall rather than familiarity
- Avoid restricting learner misconceptions
- Makes guessing and cheating harder
- Are closer in style to the questions learners face in examinations

Our platform supports free-text questions using a rule-based marking system

Incorporated [Free-Text marking component](#) of The Open University's open-source project, [OpenMark](#).

Rule	Response
<p>Value</p> <input type="text" value="gravity gravitational weight normal reaction contact surface floor table"/>  	<p>Feedback: </p> <div style="border: 1px solid #ccc; padding: 5px; width: fit-content;"><p>Correct! 😊</p></div>
<p>Ignore case ✓ Any order ✓ Extra words ✓ Misspelling ✓</p>	

Marking rules consist of:

- Word level modifiers (Regular Expression style wildcards)
- Phrase level modifiers (i.e. allow out-of-order, missing words, etc.)
- Can be combined using logical operators

We collaborated with the OU to assess how the use of free-text questions affect learners

Collaboration between Isaac and [Dr Holly Hedgeland](#) and [Mark Parker](#), of The Open University, into the effect of asking questions based on the [Force Concept Inventory](#) as free-text questions rather than multiple choice questions.

Rules were built from an initial sample of about 100 attempts at each question.

Questions were hosted on Isaac Physics with nearly 2,000 question attempts over the 32 questions.

Accuracy was evaluated and iterated upon to improve future marking accuracy.

The rule-based system marks sufficiently accurately but rule creation can be expensive

Rule-based marking can achieve high accuracy on unseen data, which is comparable to human marking.

Marking rules, in general, are constructed from an initial set of about 60 human marked questions and an average accuracy of around 0.96 has been reported in the [literature](#).

Question creation requires effort and skill.

Question ID	Accuracy
ams_quiz3_4 d46dd3f2	1.00
ams_quiz1_7 684df696	1.00
ams_quiz2_6 84e49146	1.00
...	...
ams_quiz1_6 7fc744eb	0.84
ams_quiz3_2 db6177f5	0.84
ams_quiz3_2 ab9d4220	0.75
Average	0.94

We created a tool to help free-text question authoring

Key Features:

- [Open-source](#)
- Test harness
- JSON & CSV import
- Synonym suggestions... (coming soon!)

[DEMO]

Free-text question builder

Matching rules

Rule	Response
Value	Feedback: ✕
<input type="text" value="gravity gravitational weight"/>	<div>Correct ! 😊</div>
Ignore case ✓	
Any order ✓	
Extra words ✓	
Misspelling ✓	

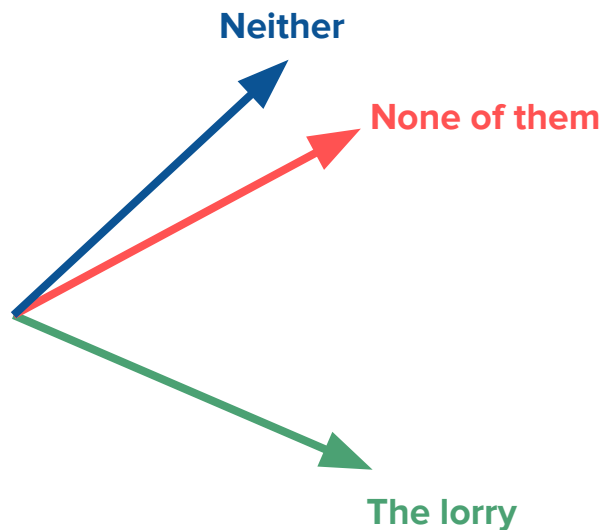
+

Test answers (1/1)

Expected	Value	Actual	Feedback	Match
✓	<input type="text" value="Gravity is the reason"/>	✓	Correct! 😊	✓ ✕

Can we do any
better?

The state-of-the-art in NLP has recently made noticeable advancements



Distance	Neither	None of them	The lorry
Neither	0	0.05	0.65
None of them	0.05	0	0.68
The lorry	0.65	0.68	0

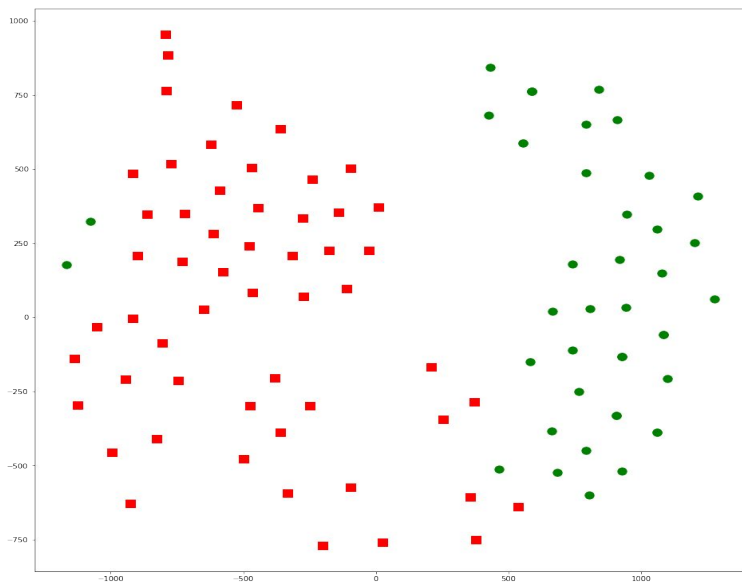
These pretrained models can be used to create answer embeddings

By using [Sentence-BERT](#), a modification of the pre-trained [BERT](#) network, for quick sentence similarity comparison, student answers can be fed into the pre-trained model to produce each answer's n-dimensional embedding.

u correct answers and v incorrect answers are selected as model answers.

Unseen test answers are assigned the human mark of the closest model answer through cosine similarity (nearest neighbour classification).

Answer embeddings seem to cluster around semantically similar responses



BERT sentence embedding t-SNE plot

After dimensionality reduction from 700+ to 2 in the t-SNE plots, clustering of correct (green) and incorrect (red) is still prominent.

Cluster analysis could be used to detect common wrong answers and potentially categories of misconceptions.

A pre-test shows that high accuracy is achievable, potentially even with only a few training cases

Answers sampled randomly to become our correct and incorrect model answers.

Test answers are assigned mark of nearest model answer.

Average accuracies reported over 100 runs.

Question ID	Rule Based Accuracy	SE (C: 10, F: 20) Accuracy	SE (C: 5, F: 5) Accuracy	SE (C: 1, F: 4) Accuracy
Ams_quiz1_1_a	0.96	0.96	0.95	0.79
ams_quiz1_2_a	0.97	0.99	0.96	0.90
ams_quiz1_2_b	0.96	0.86	0.82	0.59
ams_quiz1_3_a	0.99	0.88	0.74	0.82
...
ams_quiz3_4_a	1.00	1.00	1.00	0.98
ams_quiz3_5_a	0.98	1.00	1.00	1.00
ams_quiz3_5_b	0.97	1.00	0.94	0.94
ams_quiz3_7_a	0.95	0.97	0.90	0.56
Average	0.94	0.93	0.88	0.76

Answer embeddings could:

Simplify question construction

Require less training samples

Cluster for common wrong answers

These pre-test results should encourage further research

- Test the effectiveness of converting multiple-choice questions into free-text questions
- Check to see how accurately cluster analysis on sentence embeddings can categorise different types of misconceptions
- BERT models can be fine-tuned
- Language models could be used for semantic paraphrasing to boost the number of initial samples

Considerations

- The effect of false positive and false negative results on students
- Unintended biases

Thank you for listening!

Please get in touch with questions or enquiries

mlt47@cam.ac.uk

