EXPLAINABILITY IN AUTOMATIC SHORT ANSWER GRADING
AGENDA

Introduction

Related Work

Explainability in Automatic Short Answer Grading

User Study

Conclusion and Future Work
MOTIVATION: UN Sustainable Development Goal 4

MOTIVATION: Challenges in Education

ACCELERATION

COSTS

SCALING
AI IN EDUCATION: Potential

AUTOMATION

GRADING

Image Source: OpenClipart-Vectors/154119/Pixabay.
### Question
What is a variable?

#### Model answer
A location in memory that can store a value.

#### Example: Answer 1
A variable is a location in memory where a value can be stored.

#### Grading: Answer 1
5 of 5 points

#### Example: Answer 2
Variable can be an integer or a string in a program.

#### Grading: Answer 2
2 of 5 points
AUTO-GRADING: Trust

STUDY

− 70 professors, lecturers and teachers

EXPLAINABILITY IN AUTOMATIC SHORT ANSWER GRADING. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.
RELATED WORK
RELATED WORK: AUTO-GRADING

AUTOMATIC SHORT ANSWER GRADING

Deep learning

e.g., (Burrows et al., 2014; Camus & Filighera, 2020; Sawatzki et al., 2021; Schlippe & Sawatzki, 2021b)
RELATED WORK: CROSS-LINGUAL AUTO-GRADING

AUTOMATIC SHORT ANSWER GRADING

Deep learning

Cross-lingual Automatic Short Answer Grading

Graphic Source: Custom Depiction.
### RELATED WORK: AUTO-GRADING

<table>
<thead>
<tr>
<th></th>
<th>multi+</th>
<th>multi+</th>
<th>multi+</th>
<th>multi+</th>
<th>multi+</th>
<th>multi+</th>
<th>mono</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en</td>
<td>de</td>
<td>nl</td>
<td>jp</td>
<td>zh</td>
<td>fi</td>
<td>6</td>
</tr>
<tr>
<td>en</td>
<td>0.45</td>
<td>0.61</td>
<td>0.64</td>
<td>0.68</td>
<td>0.63</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>ceb</td>
<td>0.70</td>
<td>0.73</td>
<td>0.72</td>
<td>0.68</td>
<td>0.72</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>sv</td>
<td>0.63</td>
<td>0.67</td>
<td>0.68</td>
<td>0.73</td>
<td>0.72</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>0.64</td>
<td>0.51</td>
<td>0.67</td>
<td>0.70</td>
<td>0.70</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>0.61</td>
<td>0.66</td>
<td>0.64</td>
<td>0.67</td>
<td>0.70</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>nl</td>
<td>0.62</td>
<td>0.64</td>
<td>0.52</td>
<td>0.70</td>
<td>0.73</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>ru</td>
<td>0.68</td>
<td>0.73</td>
<td>0.83</td>
<td>0.74</td>
<td>0.75</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>it</td>
<td>0.62</td>
<td>0.65</td>
<td>0.72</td>
<td>0.71</td>
<td>0.73</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>0.61</td>
<td>0.68</td>
<td>0.76</td>
<td>0.68</td>
<td>0.72</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>pl</td>
<td>0.62</td>
<td>0.71</td>
<td>0.77</td>
<td>0.69</td>
<td>0.72</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>vi</td>
<td>0.71</td>
<td>0.72</td>
<td>0.84</td>
<td>0.77</td>
<td>0.73</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>jp</td>
<td>0.66</td>
<td>0.70</td>
<td>0.73</td>
<td>0.49</td>
<td>0.63</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>zh</td>
<td>0.63</td>
<td>0.71</td>
<td>0.77</td>
<td>0.69</td>
<td>0.50</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>ar</td>
<td>0.72</td>
<td>0.78</td>
<td>0.85</td>
<td>0.78</td>
<td>0.76</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>uk</td>
<td>0.65</td>
<td>0.70</td>
<td>0.82</td>
<td>0.73</td>
<td>0.73</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>pt</td>
<td>0.59</td>
<td>0.67</td>
<td>0.75</td>
<td>0.69</td>
<td>0.73</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>fa</td>
<td>0.64</td>
<td>0.66</td>
<td>0.71</td>
<td>0.67</td>
<td>0.70</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>ca</td>
<td>0.64</td>
<td>0.70</td>
<td>0.74</td>
<td>0.70</td>
<td>0.76</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>sr</td>
<td>0.69</td>
<td>0.81</td>
<td>0.83</td>
<td>0.76</td>
<td>0.79</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>id</td>
<td>0.66</td>
<td>0.68</td>
<td>0.69</td>
<td>0.70</td>
<td>0.79</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>0.63</td>
<td>0.69</td>
<td>0.65</td>
<td>0.75</td>
<td>0.71</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>ko</td>
<td>0.70</td>
<td>0.70</td>
<td>0.76</td>
<td>0.66</td>
<td>0.66</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>fi</td>
<td>0.69</td>
<td>0.79</td>
<td>0.77</td>
<td>0.77</td>
<td>0.73</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>hu</td>
<td>0.69</td>
<td>0.76</td>
<td>0.81</td>
<td>0.72</td>
<td>0.76</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>cs</td>
<td>0.62</td>
<td>0.77</td>
<td>0.82</td>
<td>0.72</td>
<td>0.78</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>sh</td>
<td>0.66</td>
<td>0.77</td>
<td>0.79</td>
<td>0.74</td>
<td>0.78</td>
<td>0.79</td>
<td></td>
</tr>
</tbody>
</table>

**Mean Absolute Error**

- 0.75 points
- Human grader variability

Graphic Sources: Schlippe & Sawatzki (2021b).

Mean Absolute Error out of 5 points
RELATED WORK: EXPLAINABILITY

Explainability in Automatic Short Answer Grading. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.

Source: Barredo Arrieta et al. (2020).
RELATED WORK: EXPLAINABILITY

Explainability in Automatic Short Answer Grading. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AET 2022.

ADDITIONAL PREDICTIVE MODELS that can meet the accuracy of black box models while being intrinsically interpretable.
## RELATED WORK: EXPLAINABILITY

<table>
<thead>
<tr>
<th>XAI method class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>confidence score</td>
<td>Certainty of a model’s prediction is made interpretable and inspectable</td>
</tr>
<tr>
<td></td>
<td>(van der Waa et al., 2020)</td>
</tr>
<tr>
<td>word highlighting</td>
<td>Words are color marked to indicate their relevance towards the classification</td>
</tr>
<tr>
<td></td>
<td>(Ribeiro, Singh &amp; Guestrin, 2016)</td>
</tr>
<tr>
<td>concept activation</td>
<td>High level human concepts are used to explain a classification</td>
</tr>
<tr>
<td></td>
<td>(Kim at al., 2018)</td>
</tr>
</tbody>
</table>
## RELATED WORK: EXPLAINABILITY

<table>
<thead>
<tr>
<th>XAI method class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>confidence score</td>
<td>Certainty of a model’s prediction is made interpretable and inspectable (van der Waa et al., 2020)</td>
</tr>
<tr>
<td>word highlighting</td>
<td>Words are color marked to indicate their relevance towards the classification (Ribeiro, Singh &amp; Guestrin, 2016)</td>
</tr>
<tr>
<td>concept activation</td>
<td>High level human concepts are used to explain a classification (Kim et al., 2018)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>XAI method for ASAG</th>
<th>XAI method class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Points</td>
<td></td>
</tr>
<tr>
<td>Predicted Points with Confidence Scores</td>
<td>confidence score</td>
</tr>
<tr>
<td>Predicted Points with Confidence Scores and Similar Answers</td>
<td>confidence score</td>
</tr>
<tr>
<td>Predicted Points with Relevance of Words in the Answer</td>
<td>word highlighting</td>
</tr>
<tr>
<td>Predicted Points with Matching Positions</td>
<td>concept activation</td>
</tr>
</tbody>
</table>
EXPLAINABILITY IN AUTOMATIC SHORT ANSWER GRADING
AUTO-GRADING: Explainability: XAI methods

AI PREDICTION

2 points

points
AUTO-GRADING: Explainability: XAI methods

EXPLANATION
11 of 12 comparable answers were also rated with 2 points by humans. That corresponds to 92%.

confidence
**AUTO-GRADING: Explainability: XAI methods**

<table>
<thead>
<tr>
<th>AI PREDICTION</th>
<th>CONFIDENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 points</td>
<td>LOW 83%</td>
</tr>
<tr>
<td>HIGH</td>
<td></td>
</tr>
</tbody>
</table>

**EXPLANATION**

10 of 12 comparable answers were also rated with 2 points by humans. That corresponds to 83%.

**EXAMPLE OF SIMILAR ANSWERS**

An animal that lowers or raises its temperature depending on its environment.

**Score:** 2 points

*confidence + similar answers*
AUTO-GRADING: Explainability: XAI methods

Explainability in Automatic Short Answer Grading. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.

relevance of words
## AUTO-GRADING: Explainability: XAI methods

<table>
<thead>
<tr>
<th>AI PREDICTION</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 points</td>
<td>The AI assistant compared with the model answer and found 2 content matches.</td>
</tr>
<tr>
<td></td>
<td>Everything we consciously perceive[\text{+1}] such as vision, is processed and transmitted through body parts[\text{+1}].</td>
</tr>
<tr>
<td></td>
<td>points + matching positions</td>
</tr>
</tbody>
</table>
AUTO-GRADING: Explainability

STUDY

— 5 XAI methods

— 70 professors, lecturers and teachers

— 9 aspects evaluated

✓ trust, ✓ informative content, ✓ speed, ✓ consistency & fairness, ✓ fun,
✓ comprehensibility, ✓ applicability, ✓ use in exam preparation, ✓ in general
AUTO-GRADING: Explainability: XAI methods

AI PREDICTION

2 points

CONFIDENCE

LOW 92% HIGH

EXPLANATION

11 of 12 comparable answers were also rated with 2 points by humans. That corresponds to 92%.

EXPLANATION

10 of 12 comparable answers were also rated with 2 points by humans. That corresponds to 83%.

EXAMPLE OF SIMILAR ANSWERS

An animal that lowers or raises its temperature depending on its environment.

Score: 2 points

AI PREDICTION

2 points

EXPLANATION

The AI assistant compared with the model answer and found 2 content matches.

A kind of precipitation consisting of small pieces of ice.

Irrelevant Very relevant

Relevance of words

✓ trust, ✓ informative content, ✓ speed, ✓ consistency & fairness, ✓ fun,
✓ comprehensibility, ✓ applicability, ✓ use in exam preparation, ✓ in general

AI PREDICTION

2 points

EXPLANATION

Highlighted in color, you can see how relevant the individual words were for the AI.

We consciously perceive such as vision, processed and transmitted through body parts.
AUTO-GRADING: Explainability: XAI methods

What do you think of the concepts?

![Chart showing the points, relevance of words, points + matching positions, confidence, and confidence + similar answers for different levels of grading (very poor to very good).]

- Points
- Relevance of words
- Points + matching positions
- Confidence
- Confidence + similar answers

Average scores:
- 1: 2.54
- 2: 3.57
- 3: 3.94
- 4: 2.80
- 5: 2.97

CONCLUSION & FUTURE WORK
CONCLUSION AND FUTURE WORK

Conclusion

- Investigated and evaluated different methods for explainability in automatic short answer grading
- Survey of over 70 professors, lecturers and teachers
- Important for them to understand how the AI reaches its scoring and their confidence in an AI grading support increases when it explains itself
- Displaying the predicted points together with matches between student answer and model answer is rated better than the other tested XAI methods
- Evaluated aspects: trust, informative content, speed, consistency and fairness, fun, comprehensibility, applicability, use in exam preparation, and in general.

Future Work

- Analyze the use of our XAI methods in interactive training programs to prepare students optimally for exams
- Direct interpretation of the complex ASAG models could be also investigated
THANK YOU

Tim Schlippe
tim.schlippe@iu.org