

Fundamentos para la retroalimentación de
evaluación formativa asistida por IA para tareas
de respuesta corta en clases de inscripción grande
[Tradução do Google do Inglesa]

Matthew Beckman
Penn State University

11th International Conference on Teaching Statistics (ICOTS)

Rosário, Argentina
13 Setembro, 2022

Motivação

- A avaliação formativa beneficia alunos e instrutores (GAISE, 2016; Pearl, et al., 2012)
- Tarefas “escrever para aprender” melhoram os resultados de aprendizagem (Graham, et al., 2020)
- Crítico para os cidadãos-estatísticos comunicarem ideias estatísticas de forma eficaz (Gould, 2010)
- A prática contínua de comunicação melhora a alfabetização estatística e promove a retenção (Basu, et al., 2013)
- A colaboração homem-máquina é um mecanismo promissor para auxiliar o feedback rápido e individualizado em escala (Basu, 2013)
- O feedback assistido pela PNL foi estudado anteriormente para ensaios ou tarefas de resposta longa (ver, por exemplo, Attali, et al., 2008; Page, 1994)

Questões de pesquisa

- **RQ1:** Que nível de concordância é alcançado entre os avaliadores humanos treinados rotulando (ou seja, pontuando ou marcando) tarefas de resposta curta?
- **RQ2:** Que nível de concordância é alcançado entre avaliadores humanos e um algoritmo de PNL?
- **RQ3:** Que tipo de representação PNL leva a um bom desempenho de clustering e como isso interage com o algoritmo de classificação?

Manuscrito

Susan Lloyd, Matthew Beckman, Dennis Pearl, Rebecca Passonneau, Zhaohui Li, & Zekun Wang (accepted). Foundations of NLP-assisted formative assessment feedback for short-answer tasks in large enrollment statistics classes. Preprint URL: <http://arxiv.org/abs/2205.02829>

#Spoiler?!

- **RQ1:** Que nível de concordância é alcançado entre os avaliadores humanos treinados que rotulam (ou seja, pontuam) tarefas de resposta

Métodos (amostra)

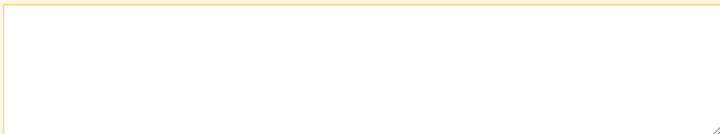
O estudo utilizou dados existentes desidentificados e rubricas de pontuação de um estudo anterior não relacionado (Beckman, 2015)

- 6 tarefas de resposta curta
- 1.935 alunos no total
- 29 turmas de 15 instituições distintas

Métodos (tarefa de resposta curta)

4. Walleye is a popular type of freshwater fish native to Canada and the Northern United States. Walleye fishing takes much more than luck; better fishermen consistently catch larger fish using knowledge about proper bait, water currents, geographic features, feeding patterns of the fish, and more. Mark and his brother Dan went on a two-week fishing trip together to determine who the better Walleye fisherman is. Each brother had his own boat and similar equipment so they could each fish in different locations and move freely throughout the area. They recorded the length of each fish that was caught during the trip, in order to find out which one of them catches larger Walleye on average.

a. Should statistical inference be used to determine whether Mark or Dan is a better Walleye fisherman? Explain why statistical inference should or should not be used in this scenario.



b. Next, explain how you would determine whether Mark or Dan is a better Walleye fisherman using the data from the fishing trip. *(Be sure to give enough detail that a classmate could easily understand your approach, and how he or she would interpret the result in the context of the problem.)*

Figure 1: Tarefa de amostra incluindo uma haste e dois prompts de resposta curta.

Métodos (Humanos)

- 3 avaliadores humanos típicos da equipe de instrução de grande número de matrículas
- respostas alocadas de tal forma que 63 respostas dos alunos em comum para cada combinação de avaliadores para quantificar a concordância
- única restrição: dados suficientes para análise intra-avaliador para pessoa que rotulou um conjunto anterior de 178 respostas 6 anos antes

Métodos (PNL)

O conjunto de respostas à tarefa foi dividido aleatoriamente de quatro maneiras:

- 90% foram divididos na divisão típica de treinamento (72%), desenvolvimento (9%) e teste (9%)
- 10% mantidos em reserva para testes mais rigorosos

Dois algoritmos de PNL foram comparados quanto à precisão usando um subconjunto de respostas dos alunos (Li et al., 2021).

- LSTM: uma regressão logística combinada com uma Long Short-Term Memory para aprender representações vetoriais
- SFRN: Rede de relação de transformação semântica baseada em recursos

Resultados (RQ1)

RQ1: Que nível de concordância é alcançado entre avaliadores humanos treinados que rotulam (ou seja, pontuam) tarefas de resposta curta?

Comparação	Confiabilidade
Avaliador A e Avaliador C	QWK = 0,83
Avaliador A e Avaliador D	QWK = 0,80
Avaliador C & Avaliador D	QWK = 0,79
Avaliador A: 2015 e 2021	QWK = 0,88
Avaliadores A, C e D	FK = 0,70

Interpretação de confiabilidade¹: $0,6 < \text{substancial} < 0,8 < \text{quase perfeito} < 1,0$

¹Viera & Garrett (2005)

Resultados (RQ2)

RQ2: Que nível de concordância é alcançado entre avaliadores humanos e um algoritmo de PNL?

O algoritmo SFRN alcançou uma precisão de classificação muito maior do que o LSTM (83% vs. 72%)². Acordo Humano e SFRN:

Comparação	Confiabilidade
Avaliador A & SFRN	QWK = 0,79
Avaliador C & SFRN	QWK = 0,82
Avaliador D & SFRN	QWK = 0,74
Avaliadores: A, C, D e SFRN	FK = 0,68

Interpretação de confiabilidade³: $0,6 < \text{substancial} < 0,8 < \text{quase perfeito} < 1,0$

²a comparação SFRN e LSTM exclui instâncias em que os rótulos humanos discordam

³Viera & Garrett (2005)

Resultados (RQ3)

RQ3: Que tipo de representação de PNL leva a um bom desempenho de clustering e como isso interage com o algoritmo de classificação?

- SFRN aprende uma representação vetorial de alta dimensão ($D = 512$) nos dados de treinamento.
- Experimentos com agrupamento K-means e K-medoids mostraram que o SFRN produz agrupamentos mais consistentes quando treinado novamente (0,62), em comparação com outros classificadores.⁴
- Maior consistência (0,88; $D = 50$), no entanto, foi alcançada usando um método de fatoração de matrizes que produz representações estáticas (WTMF; Guo & Diab, 2011)

⁴A consistência é medida como a proporção de todos os pares de respostas em uma determinada classe por pergunta que são agrupadas da mesma maneira em duas execuções (no mesmo agrupamento ou não no mesmo agrupamento).

Discussão

- **RQ1:** A concordância substancial alcançada entre avaliadores humanos treinados fornece contexto para comparações adicionais
- **RQ2:** algoritmo de PNL produziu concordância razoavelmente alinhada aos resultados alcançados por pares/grupos de avaliadores humanos treinados
- **RQ3:** Classificação e agrupamento têm incentivos concorrentes para dimensionalidade; Low D é melhor para a estabilidade do cluster, High D é melhor para a confiabilidade da classificação.

Trabajo futuro

- El estudio utiliza datos existentes de estudios anteriores recopilados de muchas clases de diferentes tamaños
 - ni una sola clase grande
 - esperamos que los resultados observados sean conservadores debido a la variabilidad adicional entre instituciones e instructores, pero se investigarán más a fondo
- “Maldición de la dimensionalidad” en el lado del aprendizaje automático
- Rendimiento de agrupamiento frente a significado semántico
 - el agrupamiento es necesario, pero no suficiente, para el significado semántico
 - el significado semántico de los grupos de PNL aún no se ha estudiado rigurosamente

Referências (1/2)

- 1 Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). Automated Scoring of Short-Answer Open-Ended Gre® Subject Test Items. *ETS Research Report Series, 2008*(1), i–22.
- 2 Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402. https://doi.org/10.1162/tacl_a_00236
- 3 Beckman, M. (2015). Assessment Of Cognitive Transfer Outcomes For Students Of Introductory Statistics.
<http://conservancy.umn.edu/handle/11299/175709>
- 4 GAISE College Report ASA Revision Committee (2016). Guidelines for Assessment and Instruction in Statistics Education College Report 2016. URL: <http://www.amstat.org/education/gaise>
- 5 Gould, R. (2010). Statistics and the Modern Student. *International Statistical Review / Revue Internationale de Statistique, 78*(2), 297–315. <https://www.jstor.org/stable/27919839>
- 6 Guo, W., Diab, M. (2012) Modeling Sentences in the Latent Space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 864–872. Association for Computational Linguistics.

Referências (2/2)

- 7 Graham, S., Kiuahara, S. A., & MacKay, M. (2020). The Effects of Writing on Learning in Science, Social Studies, and Mathematics: A Meta-Analysis. *Review of Educational Research*, 90(2), 179–226. <https://doi.org/10.3102/0034654320914744>
- 8 Li, Z., Tomar, Y., & Passonneau, R. J. (2021). A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6030–6040. Association for Computational Linguistics. <https://aclanthology.org/2021.emnlp-main.487>
- 9 Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *The Journal of Experimental Education*, 62(2), 127–142.
- 10 Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics. URL: http://www.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf
- 11 Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5), 360–363.

Obrigada

Fundamentos para la retroalimentación de
evaluación formativa asistida por IA para tareas
de respuesta corta en clases de inscripción grande

[Tradução do Google do Inglesa]

Matthew Beckman
Penn State University

11th International Conference on Teaching Statistics (ICOTS)

Rosário, Argentina
13 Setembro, 2022