Foundations for scalable NLP-assisted formative assessment feedback

Matthew Beckman
Penn State University

April 3, 2023

Foundations for scalable NLP-assisted formative assessment feedback

Matthew Beckman
Penn State University

April 3, 2023

Two question survey before seminar (scan with mobile phone)

Figure 1: (QR Code) https://forms.gle/hpW72fMYE1SsB19JA

How did you respond to our survey?

1. Is your lucky number odd or even?
2. How did you describe the value of formative assessment?

Motivation

- "Write-to-learn" tasks improve learning outcomes (Graham, et al., 2020)
- Critical for citizen-statisticians to communicate statistical ideas effectively (Gould, 2010)
- Continual practice with communicating improves statistical literacy and promotes retention (Basu, et al., 2013)
- Formative assessment benefits both students & instructors (GAISE, 2016; Pearl, et al., 2012)
- *Logistics* of constructed response tasks jeopardize use in large-enrollment classes

Goal state

Computer-assisted formative assessment feedback for short-answer tasks in large-enrollment classes, such that instructor burden is similar to small class (~30 students)

- Human-machine collaboration is a promising mechanism to assist rapid, individualized feedback at scale (Basu, 2013)
- Natural language processing (NLP) involves how computers can be programmed to analyze language elements (e.g., text or speech)
- NLP-assisted feedback has previously been studied for essays or long-answer tasks (see e.g., Attali, et al., 2008; Page, 1994)

Research Questions

- **RQ1**: What level of agreement is achieved among trained human raters labeling (i.e., scoring or marking) short-answer tasks?
- **RQ2**: What level of agreement is achieved between human raters and an NLP algorithm?
- **RQ3**: What sort of NLP representation leads to good clustering performance, and how does that interact with the classification algorithm?

Preprint

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

Spoilers?!

- **RQ1**: What level of agreement is achieved among trained human raters labeling (i.e., scoring) short-answer tasks?
- **RQ2**: What level of agreement is achieved between human raters and an NLP algorithm?
- **RQ3**: What sort of NLP representation leads to good clustering performance, and how does that interact with the classification algorithm?

Spoilers?!

- RQ1: substantial inter-rater & intra-rater agreement
- RQ2: substantial agreement among human & NLP labeling
- RQ3: in progress, but promising

Methods (Sample)

Study utilized de-identified extant data & scoring rubrics (Beckman, 2015)

- 6 short-answer tasks
- 1,935 students total
- 29 class sections 15 distinct institutions

Methods (Short-answer task)

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 2: Sample task including a stem and two short-answer prompts.

Methods (RQ1)

- 3 human raters typical of large-enrollment instruction team
- 63 student responses in common for each *combination* of raters to quantify agreement (e.g., pairwise, consensus, etc)
- constraint: sufficient data for intra-rater analysis for person that had labeled 178 responses 6 years prior

Methods (RQ2)

The set of task-responses were randomly split four ways:

- 90% of data for development purposes, were partitioned according to machine-learning best practice:
  - training (72%),
  - development (9%)
  - evaluation (9%)
- 10% of data being held in reserve for more rigorous testing

Two NLP algorithms were compared for accuracy using a subset of student responses (Li et al., 2021).

- LSTM: a logistic regression combined with a Long Short-Term Memory for learning vector representations
- SFRN: Semantic Feature-Wise Transformation Relation Network

Methods (RQ3)

Manual pilot of human-generated clustering

- Two reviewers independently evaluated 100 student responses that earned "partial credit" on inference tasks
- Each reviewer provided free-text feedback to each student
- Verbatim feedback captured for each reviewer and cross-tabulated for analysis.

Results (RQ1)

RQ1: What level of agreement is achieved among trained human raters labeling (i.e., scoring) short-answer tasks?

Comparison	Reliability
Rater A & Rater C	QWK $= 0.83$
Rater A & Rater D	QWK $= 0.80$
Rater C & Rater D	QWK $= 0.79$
Rater A: 2015 & 2021	QWK $= 0.88$
Raters A, C, & D	FK $= 0.70$

Reliability interpretation[1]: $0.6 <$ substantial $< 0.8 <$ near perfect < 1.0

[1]Viera & Garrett (2005)

Results (RQ2)

RQ2: What level of agreement is achieved between human raters and the machine (an NLP algorithm)?

The SFRN algorithm achieved much higher classification accuracy than LSTM (83% vs. 72%)[2]. Human & SFRN agreement:

Comparison	Reliability
Rater A & SFRN	QWK $= 0.79$
Rater C & SFRN	QWK $= 0.82$
Rater D & SFRN	QWK $= 0.74$
Raters: A, C, D, & SFRN	FK $= 0.68$

Reliability interpretation[3]: $0.6 <$ substantial $< 0.8 <$ near perfect < 1.0

[2] SFRN & LSTM comparison excludes instances when human labels disagree
[3] Viera & Garrett (2005)

Results (RQ3 machines)

RQ3: What sort of NLP representation leads to good clustering performance, and how does that interact with the classification algorithm?

- SFRN learns a high-dimension (D = 512) vector representation on training data.
- Experiments with K-means and K-medoids clustering showed SFRN produce more consistent clusters when retrained (0.62), in comparison to other classifiers.[4]
- Highest consistency (0.88; D = 50), however, was achieved using a matrix factorization method that produces static representations (WTMF; Guo & Diab, 2011)

[4] Consistency is measured as the ratio of all pairs of responses in a given class per question that are clustered the same way on two runs (in the same cluster, or not in the same cluster).

Results (RQ3 humans)

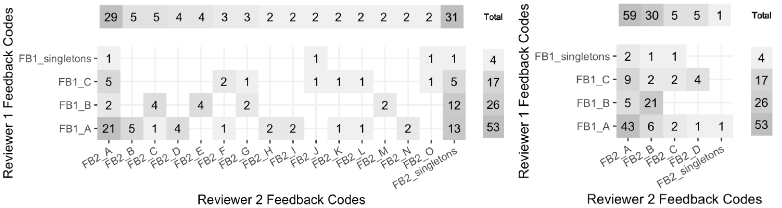

Figure 3: Cross-tabulation of feedback distribution for the two reviewers for the initial feedback (left) compared with the same analysis for the portion of feedback related to the statistical concept at issue (right).

- Reviewer 1 favored feedback on statistical concepts (only).
- Reviewer 2 provided same, plus a quote from the student
- Reviewer 2 parsed her feedback to compare her remarks related to the statistical concepts (only) with the feedback of Reviewer 1.

Discussion

- **RQ1**: Substantial agreement achieved among trained human raters provides context for further comparisons
- **RQ2**: NLP algorithm produced agreement reasonably aligned to results achieved by pairs/groups of trained human raters
- **RQ3**: Classification and clustering have competing incentives for dimensionality; Low D is better for cluster stability, High D better for classification reliability.

Future work

- Study uses extant data from prior study collected from many classes of varying size
    - not a single large class
    - we expect observed results are conservative due to additional variability across institutions and instructors, but will be investigated further
- "Curse of dimensionality" on the machine learning side
- Clustering performance vs semantic meaning
    - clustering is necessary, but not sufficient, for semantic meaning
    - semantic meaning of NLP clusters not yet rigorously studied

References (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.

References (2/2)

7 Graham, S., Kiuhara, 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.

Thank You

Foundations for scalable NLP-assisted formative assessment feedback

Matthew Beckman
Penn State University

April 3, 2023

Resource Page URL:
https://mdbeckman.github.io/2023-MSU-Colloq/