Forschungsprojekt: Kinder schreiben – KI antwortet

Kurzbeschreibung

Das Forschungsprojekt untersucht die Qualität von Rückmeldungen zu Schülertexten, die von menschlichen und künstlichen Intelligenzen gegeben werden, um deren Qualität zu verbessern und dadurch Schreibkompetenzen von Schüler:innen zu fördern.

Hintergrund und Motivation

(Deutsch-)Lehrkräfte (an Gymnasien) verbringen einen erheblichen Teil ihrer Arbeitszeit mit der Korrektur von Schülertexten, das zeigen u. a. Mußmann et al. 2016 und illustriert dieses kurze Rechenbeispiel: Bei einer Klasse mit 25 Schüler:innen, die 4 Klassenarbeiten im Schuljahr und einen vorbereitenden Übungsaufsatz schreibt, korrigiert eine Lehrkraft mit 3 Deutschklassen insgesamt 600 Aufsätze pro Schuljahr. Das sind ca. zwei Aufsätze pro Tag in nur einem Fach.

Trotz des hohen Aufwands ist die Qualität der menschlichen Rückmeldungen oft heterogen und zeitlich verzögert, was nicht immer lernförderlich ist (Müller et al. 2023). Der aktuelle Lehrkräftemangel (KMK 2023) verschärft das Problem zusätzlich. Generative Sprachmodelle, wie sie in neueren KI-Anwendungen verwendet werden, könnten theoretisch Lösungen bieten, um Lehrkräfte zu entlasten (Kasneci et al. 2023). Allerdings ist über die Qualität dieser maschinellen Rückmeldungen für den deutschsprachigen Raum bisher so gut wie nichts bekannt, insbesondere vor dem Hintergrund, dass es sich bei großen Sprachmodellen lediglich um „Sprachgebrauchsautomaten“ (Müller/Fürstenberg 2023: 1) handelt, die „nichts verstehen“ (Fürstenberg/Müller 2024: 1).

Begriffsklärung

Mit *Rückmeldungen* zu Schülertexten sind folgende Formen gemeint:

	Medialität	
mündlich	☐ ☐ ☐ ☐ ☐ ■	schriftlich
	Professionalisierung	
laienhaft	☐ ☐ ☐ ☐ ☐ ■	professionell
	Bekanntheit des Textes	
fremd	■ ☐ ☐ ☐ ☐ ☐	eigen
	Frequenz	
einfach	☐ ■ ☐ ☐ ☐ ☐	mehrfach
	Form	
analog	☐ ☐ ☐ ☐ ☐ ■	digital
	Art	
qualitativ	☐ ☐ ■ ☐ ☐ ☐	quantitativ
	Grundlage	
kriterial	■ ☐ ☐ ☐ ☐ ☐	holistisch
	Ziel	
fördern	■ ☐ ☐ ☐ ☐ ☐	benoten

Dr. phil. Maurice Fürstenberg
Universität Tübingen · Deutsches Seminar
maurice.fuerstenberg@uni-tuebingen.de

POSTDOC-KOLLEG

Es geht also um digitale, schriftliche Rückmeldungen durch Expert:innen, die einen fremden Text betreffen. Die Rückmeldungen sollen eher einmal gegeben werden und es geht eher um qualitative Rückmeldungen in Textform (z. B. Lehrerkommentar). Die Rückmeldungen kommen dabei kriterial zustande und sollen fördern. Es werden im Projekt sowohl menschliche als auch durch maschinelle Rückmeldungen analysiert.

Zur Qualität maschineller Rückmeldungen zu Schülertexten

Die maschinelle Bewertung von Schülertexten ist insbesondere für den deutschsprachigen Raum noch nicht gut untersucht, obwohl sie in Form von Automated Essay Scoring (AES) bereits in den 1960er-Jahren eingeführt (Page 1966) wurde. Die Rückmeldungen konzentrierten sich bis in die 2020er-Jahre hauptsächlich auf die Beurteilung sprachlicher Oberflächenmerkmale (einen Überblick geben Ramesh/Sanampudi 2022) und in der Folge vor allem auf einzelne Sätze bzw. sehr kurze, englische Texte.

- Die Studie von Sawatzki et al. (2022) verglich beispielweise ein auf Bidirectional Encoder Representations from Transformers (BERT) basierendes Modell mit menschlichen Bewertungen von Kurzantworten. Die Bewertungen stammten von Antworten, die BWL-Studierende in einem Moodle-Kurs gegeben hatten. Das Modell, das ‚nur' auf Wikipedia und Open Legal Data trainiert wurde, erreichte eine Korrelation von r=0.78, allerdings wird kein p-Wert angegeben, was das Ergebnis statistisch ungesichert zurücklässt.
- Padò et al. (2023) untersuchten die Inter-Rater-Reliabilität (IRR) von S(entence)-BERT-Modellen im Vergleich zu menschlichen Bewertungen im Short Answer Grading (SAG). Trotz einer Minimalskalierung auf dichotome Bewertungen (Ja/Nein) lag die mittlere Genauigkeit der SBERT-Modelle nur bei 71,4 % (MIN=64,7 %, MAX=86,3 %).
- Zahlreiche Studien befassen sich mit dem Einsatz von Large Language Models (LLMs) zur Rückmeldung auf englische Texte, darunter die Arbeiten von Mizumoto/Eguchi (2023), Ouyang et al. (2022), Naismith et al. (2023) und Chiang/Lee (2023). Letztere verglichen Bewertungen von 400 englischen Textfragmenten (zur Hälfte von Menschen, zur Hälfte von GPT-2 verfasst) durch GPT-3 und drei bezahlte Lehrkräfte. Die Texte wurden in vier Kategorien (*Grammaticality*, *Cohesion*, *Likability*, *Relevance*) auf einer 5-stufigen Likert-Skala bewertet. Nur für eine Kategorie (*Relevance*) ergab sich ein starker Zusammenhang (τ=0.38).
- Eine aktuelle Studie von Stahl et al. (2024) prüfte verschiedene Prompting-Strategien für zero-shot und little-shot learning, um zu untersuchen, wie gut Mistral Rückmeldungen zu englischen Texten gibt. Hierbei wurden die Rückmeldungen von Mistral mit denen von LLaMA-2 und menschlichen Ratern (Laien, N=12) verglichen. Die Autor:innen rechtfertigen die Wahl, ein Modell von einem anderen bewerten zu lassen, mit Verweis auf Chiang/Lee (2023): „Using LLMs to assess the quality of generated texts has been shown to be consistent with human expert annotations for some free-text generation tasks." Für diese generale Aussage reichen die Ergebnisse von Chiang/Lee (2023) allerdings bei weitem nicht aus.

Obwohl viele Studien verschiedene Metriken zur Bewertung von Texten untersuchen, gibt es nur wenige, die sich auf deutsche Texte konzentrieren. Authentische Lehrerurteile zu echten Schülertexten werden selten untersucht und Studien von Deutschdidaktiker:innen sind praktisch nicht vorhanden.

Dr. phil. Maurice Fürstenberg
Universität Tübingen · Deutsches Seminar
maurice.fuerstenberg@uni-tuebingen.de

Untersuchungsdesign

Das Forschungsdesign integriert sowohl quantitative als auch qualitative und gemischte Ansätze:

1. **Quantitative Analyse**: Vergleich kriteriengeleiteter, quantifizierter Urteile durch menschliche Lehrkräfte und KI.
2. **Qualitative Analyse**: Expert:innen bewerten die Qualität der Rückmeldungen von generativer KI.
3. **Mixed-Method**: Fragebogenuntersuchung zur Meinung von Schüler:innen zu menschlichem und maschinellem Feedback.

Zusätzlich soll untersucht werden, wie sich die Qualität der maschinellen Rückmeldungen steigern lässt. Hierbei wären grundsätzlich zwei Ansätze denkbar:

- **Bottom-up-Training**: Durch ein großes Korpus aus Schülertexten, ergänzt durch Noten und Lehrerkommentare soll ein Sprachmodell in die Lage versetzt werden, typische Muster von professionellen Rückmeldungen zu erkennen und lernförderliche Rückmeldungen zu geben.
- **Top-down-Prompting**: Eine gezielte Vorgabe von konkreten Beurteilungskriterien soll die Rückmeldung der KI verbessern.

Da für das Bottom-up-Training eine riesige Menge an bisher nicht für die Forschung zugänglichen Daten notwendig ist, wird dieser Ansatz zurückgestellt und durch das Top-down-Prompting ersetzt. Dafür sind jedoch möglichst konkrete Beurteilungskriterien zu konkreten (didaktischen) Textsorten, für bestimmte Jahrgangsstufen und Bildungsgänge. Dies stellt für die meisten Textsorten ein Desiderat dar, das ebenfalls untersucht wird. Zwar haben sich Beurteilungskriterien schon einige Projekte gewidmet (u. a. Sturm 2016, Sieber 2019, Vanselow et al. 2022) und vereinzelt sind auch textsortenspezifische Kriterien vorhanden (u. a. Pissarek 2010, Rödel 2016, Wild 2022), allerdings sind diese für Lehrkräfte oft wenig zugänglich.

Ausblick

Langfristig soll ein konsistentes Set an Bewertungskriterien entwickelt werden, das von Lehrkräften und Didaktiker:innen in Workshops erarbeitet wird. Diese Kriterien sollen auf unterschiedliche Textsorten und Altersstufen anwendbar sein und die Grundlage für ein effektives Feedback darstellen. Das Projekt strebt an, einen Common Sense für textsortenspezifische und prozessorientierte Rückmeldungen zu schaffen.

Geplant sind bereits erste Workshops mit Lehrkräften, um die entwickelten Kriterien zu verfeinern und die KI-basierten Rückmeldungen weiter zu testen. Auch *fiete.ai* (Haverkamp/Hecht/Schindler 2024) hilft durch die Freigabe von Bewertungskriterien, die Lehrkräfte auf dieser Plattform angeben. Parallel dazu werden und wurden erste Studien zur Qualität maschineller Rückmeldungen durchgeführt. Langfristig wird eine Förderung durch die Deutsche Forschungsgemeinschaft (DFG) und die Einbindung von Doktorand:innen angestrebt, um angestrebten flächenwirksam untersuchen zu können.

Dr. phil. Maurice Fürstenberg
Universität Tübingen · Deutsches Seminar
maurice.fuerstenberg@uni-tuebingen.de

Literatur

Chiang, C.-H./Lee, H. (2023). Can large language models be an alternative to human evaluations? In: Rogers, A., Boyd-Graber, J., Okazaki, N. (Hg.): *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* S. 15607–15631.

Fürstenberg, M./Müller, H.-G. (2024): KI im Deutschunterricht. In: *Der Deutschunterricht*, Heft 5, S. 2–13.

Haverkamp, H./Hecht, M./Schindler, K. (2024): Lernförderliches Feedback KI-basiert vermitteln. In: *Der Deutschunterricht*, Heft 5, S. 60–71.

Kasneci, E./Seßler, K./Küchemann, S./Bannert, M./Dementieva, D./Fischer, F./Gasser, U./Groh, G./Günnemann, S./Hüllermeier, E. et al. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. In: *Learning and individual differences*, 103: 102274.

Mizumoto, A./Eguchi, M. (2023). Exploring the potential of using an ai language model for automated essay scoring. In: *Research Methods in Applied Linguistics*, 2(2):100050.

Mußmann, F./Riethmüller, M./Hardwig, T./Peters, S./Parciak, M./Ohms, I. C./Klötzer, S. (2016). Niedersächsische Arbeitszeitstudie 2015 / 2016: Lehrkräfte an öffentlichen Schulen. Ergebnisbericht. Online unter: https://www.gew-nds.de/fileadmin/media/sonstige_downloads/nds/Mehrarbeit/Niedersaechsische-Arbeitszeitstudie2015-2016-Endbericht.pdf (13.07.2024).

Müller, H.-G./Fürstenberg, M. (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Deutschen Germanistenverbandes* Jg. 70, Heft 4, S. 327–345.

Müller, N./Utesch, T./Busse, V. (2023) Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs- und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrationsbedingter Mehrsprachigkeit. In: *Unterrichtswissenschaft* 51, 169–198. https://doi.org/10.1007/s42010-023-00173-2

Naismith, B./Mulcaire, P./Burstein, J. (2023). Automated evaluation of written discourse coherence using gpt-4. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, S. 394–403.

Ouyang, L./Wu, J./Jiang, X./Almeida, D./Wainwright, C./Mishkin, P./Zhang, C./Agarwal, S./Slama, K./Ray, A. et al. (2022). Training language models to follow instructions with human feedback. In: *Advances in neural information processing systems*, 35: 27730–27744.

Padò, U./Eryilmaz, Y./Kirschner, L. (2023). Short-answer grading for german:Addressing the challenges. In: *International Journal of Artificial Intelligence in Education*, S. 1–32.

Page, E. B. (1966). The Imminence of... Grading Essays by Computer. The Phi Delta Kappan, 47(5), 238–243. http://www.jstor.org/stable/20371545

Pissarek, M. (2010): Und wer gewinnt? Spielanleitungen kriterienorientiert verfassen. In: Praxis Deutsch (223/37), S. 32–41.

Ramesh, D./Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. In: *Artificial Intelligence Review*, 55(3): 2495–2527.

Rödel, M. (2016): *Interpretationsaufsätze schreiben. Ein Handbuch*. Baltmannsweiler: Schneider Verlag Hohengehren.

Sawatzki, J./Schlippe, T./Benner-Wickner, M. (2022). Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In: *International conference on artificial intelligence in education technology*. Springer, 65–75.

Sieber, P. (2019): Kriterien der Textbewertung am Beispiel Parlando. In: Janich, Nina (Hg): Textlinguistik – 15 Einführungen und eine Diskussion. S. 261–280.

Stahl, M./Biermann, L./Nehring, A./Wachsmuth, H. (2024). Exploring llm prompting strategies for joint essay scoring and feedback generation. Online unter: https://arxiv.org/abs/2404.15845 (19.07.2024).

Sturm, A. (2016): Beurteilen und Kommentieren von Texten als fachdidaktisches Wissen. In: Leseräume, 3, S. 115–132.

Vanselow, L./Jansen, T./Kilian, J./Strahl, F./Möller, J. (2022): Gerecht? Zu streng? Zu mild? Das ASSET-G-Projekt zur Erforschung von Einflussfaktoren bei der Bewertung schriftlicher Leistungen im Deutschunterricht. In: *Der Deutschunterricht* 2022/4, 89–95.

Wild, J. (2022). Erzählen professionell unterrichten: Was eine Lehrkraft beim Schreiben wissen muss. forAp, 5(5), 43-62. https://doi.org/10.5283/forap.72

[Stand:19.10.2024]

Dr. phil. Maurice Fürstenberg
Universität Tübingen · Deutsches Seminar
maurice.fuerstenberg@uni-tuebingen.de