Hendrik Fleischer[1]
Conrad Borchers[2]
Sascha Schanze[1]
Vincent Aleven[2]

[1]Leibniz Universität Hannover
[2]Carnegie Mellon University Pittsburgh

# Fehlerklassifizierung beim tutor-gestützten Lösen von Stöchiometrieaufgaben

## Ausgangslage

Das Lösen von stöchiometrischen Aufgaben erfordert eine Verknüpfung mathematischer und chemischer Fertigkeiten (Marais & Combrinck, 2009). Fehler-spezifisches Feedback (Van der Kleij et al., 2015) und Hinweise stellen eine Möglichkeit dar, Lernende beim Lösen von Stöchiometrieaufgaben zu unterstützen (Aleven et al., 2016; Wang et al., 2019). In dieser Studie werden zwei Intelligente Tutorsysteme (ITS) erprobt, welche unterschiedliche Strategien beim Lösen von stöchiometrischen Aufgaben akzeptieren (Flowers & Theopold, 2019; King et al., 2022; Schmidt, 1997). Ein langfristiges Ziel ist es, die ITS dazu zu befähigen, aus Logdaten, welche jegliche Interaktionen der Nutzenden im Tutorsystem protokollieren, verschiedene Arten von Lösungsfehlern zu erkennen, um Lernende adaptiv zu unterstützen. Dieser Beitrag fokussiert die Fragestellungen, welche Fehlerklassifikationen manuell aus den durch die ITS generierten Logdaten zum Lösen von stöchiometrischen Aufgaben ausfindig gemacht und automatisch klassifiziert werden können.

## Intelligente Tutorsysteme

Im Folgenden werden der StoichTutor (McLaren et al., 2011) und der ORCCA Tutor (King et al., 2022) vorgestellt, die in den USA entwickelt wurden.

### *StoichTutor*

Im StoichTutor (s. Abb. 1) besteht das Ziel, mit gegebenen Werten (Masse oder Stoffmenge) strukturiert einen gesuchten Wert (Teilchenanzahl oder Stoffmengenkonzentration) zu berechnen (McLaren et al., 2011). Unter Berücksichtigung der *factor-label* Methode (Schmidt, 1997) müssen neben der Eingabe von Werten, ebenfalls Einheiten und Substanzen durch ein Drop-Down-Menü ausgewählt werden. Hinweise für jeden Problemlöseschritt, welche den nächsten Problemlöseschritt erklären, können angefordert werden. Weiterhin gibt der StoichTutor Feedback auf die Genauigkeit einzelner Problemlöseschritte mit konzeptuellen Hinweisen bei typischen Fehlern, wie das Vertauschen von Zähler und Nenner.

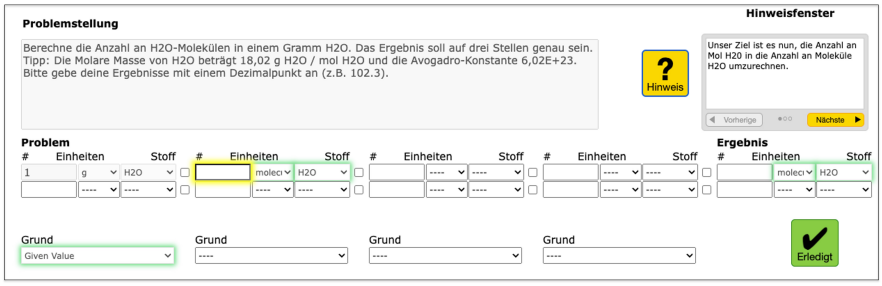


*Abb. 1 Interface des StoichTutors.*

*ORCCA*

Im ORCCA Tutor (s. Abb. 2) besteht die Möglichkeit multipler Eingabemöglichkeiten (z.B. Eingabe von Formeln), um einen gesuchten Wert zu berechnen. Der ORCCA Tutor akzeptiert mehrere Strategien, z.B. die Mol-Methode oder die Proportionalitätsmethode (Schmidt, 1997). Im Gegensatz zum StoichTutor kann die Eingabe von Substanz und Einheit im Löseprozess zusammengefasst dargestellt oder vernachlässigt werden. Lediglich das Endergebnis muss vollständig mit den Einheiten eingetragen werden. Der ORCCA Tutor erlaubt Notizen über die Tastatureingabe anzufertigen, um relevante Formeln, Zwischenwerte oder das allgemeine Vorgehen zum Lösen der Aufgaben zu notieren. Lernende erhalten generisches Feedback über die Relevanz numerischer Zwischenergebnisse sowie Hinweise auf die nächsten Problemlöseschritte.

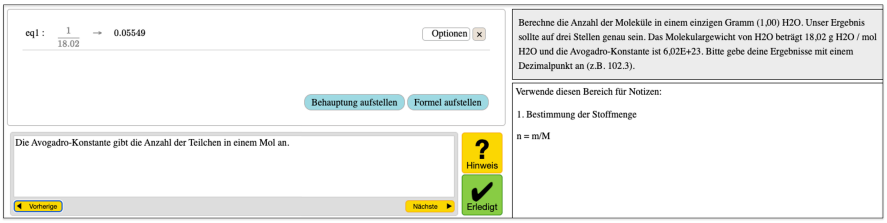


*Abb. 2 Interface des ORCCA Tutors.*

**Methodik**

Basis dieser Studie sind Logdaten von insgesamt 61 angehenden Studierenden einer Universität in Deutschland, die in beiden ITS insgesamt 14 stöchiometrische Aufgaben über drei Tagen bearbeitet haben. Zur Klassifikation der Fehlerquellen wurden s.g. Clips von Logdaten analysiert. Insgesamt wurden 478 Clips ausgewertet, welche Bereiche vom ersten inkorrekten Input bis zum ersten darauffolgenden korrekten Input umfassen. Zur Fehlerklassifikation wurde ein Kategoriensystem anhand einer expertenbasierten Auswertung der Logdaten entwickelt (s. Tab. 1). Das Kategoriensystem beruht auf den stöchiometrischen Operationen, welche zum Lösen der Stöchiometrieaufgaben notwendig sind (Gulacar et al., 2013). Aus der manuellen Kodierung wurden zwischen vier und N Entscheidungsregeln für jede Kategorie abgeleitet. Ein Beispiel für eine Entscheidungsregel für die Kategorie 4.3 Composition Stoichiometry wäre, dass wenn ein Wert um einen Multiplikator mit einer natürlichen Zahl vom Ergebnis abweicht, dann das stöchiometrische Verhältnis inkorrekt verwendet oder vernachlässigt wurde. Das Kategoriensystem fokussiert auch auf Schwierigkeiten bei der Nutzbarkeit (Usability) (Chughtai et al., 2015) sowie auf der Erkennung von Gaming the System (Paquette et al., 2014). Eine dritte Person implementierte diese Entscheidungsregeln in Python und evaluierte ihre Klassifikationsgenauigkeit anhand der kodierten Sequenzen. Zur Evaluierung der Inter-Rater Reliabilität (IRR) des Kategoriensystems haben zwei Coder die Sequenzen doppelcodiert. Die manuell doppelcodierten Sequenzen bilden nach Baker et al. (2006) die optimale Lösung (Goldstandard) zur Klassifikation von Clips und stellen folglich die Grundlage für die automatische Fehlerklassifikation dar. Bei Unstimmigkeit der beiden Coder wurde sich nach erneuter Diskussion auf eine Kategorie geeinigt.

Die Interrater-Reliabilität der manuellen Kodierung und die AUC-Werte (*area under curve*), als Gütemerkmal binärer Klassifikation individueller Fehlerkategorien sowie die Präzision der automatischen Detektion für die einzelnen Kategorien, ist Tab. 1 zu entnehmen.

*Tab 1 Ausschnitt des Kategoriensystems unter Angabe der IRR sowie des AUC der automatischen Fehlerklassifizierung basierend auf N = 478 kodierten Log Daten Clips.*

| Kategorie | IRR | AUC | Präzision | Recall |
|---|---|---|---|---|
| 1 Usability | 0.813 | 0.883 | 0.644 | 0.825 |
| 2 Interface Learning | 0.770 | 0.727 | 0.404 | 0.733 |
| 3 Gaming the system | 0.746 | 0.812 | 0.323 | 0.741 |
| 4 Chemistry Learning | | | | |
| 4.1 Unit Conversion | 0.754 | 0.673 | 0.310 | 0.466 |
| 4.2 Particle Differentiation | 0.803 | 0.818 | 0.406 | 0.722 |
| 4.3 Composition Stoichiometry | 0.679 | 0.796 | 0.278 | 0.712 |
| 4.4 $n \sim N_A$ | 0.656 | 0.757 | 0.455 | 0.526 |
| 4.5 n ~ c | 0.835 | 0.883 | 0.463 | 0.787 |
| 4.6 n ~ m | 1.000 | 0.881 | 0.619 | 0.780 |

**Vorläufige Ergebnisse**

Das Kategoriensystem weist mit einem Cohens Kappa von $\kappa$ = 0.60 - 1.00 eine gute IRR auf (Bortz et al., 2006). Gleiches gilt für die AUC mit einem Bereich von 0.60 - 0.80. Allerdings zeigt sich eine geringe Präzision der automatischen Fehlererkennung bei allen Kategorien, abgesehen von 1 Usability und 4.6 n ~ m. Kategorien mit einer auffallend geringen Präzision stellen die Kategorien 4.1 Unit Conversion und 4.3 Composition Stoichiometry dar. Die automatische Fehlerklassifikation kann die vorher manuell klassifizierten Sequenzen zwar reliabel bestätigen, doch fällt die Präzision der Kategorisierung gering aus (s. Tab. 1). Zum aktuellen Zeitpunkt werden mehr falsch-positive als falsch-negative Sequenzen klassifiziert, wobei die falsch-positive Fehlerklassifikation hier eine Klassifizierung von Fehlertypen meint, die in den Sequenzen nicht vorliegen.

**Diskussion und Ausblick**

Aus den Ergebnissen zeigt sich, dass viele der definierten Fehlertypen beim Lösen von Stöchiometrieaufgaben anhand der Logdaten durch menschliches Kodieren reliabel codiert werden können. Eine offene Forschungsfrage ist, inwiefern weitere Entscheidungsregeln generalisiert oder tutor-spezifisch zur Verbesserung der automatischen Klassifikation beitragen können. Die hohe Rate an falsch-positiv klassifizierten Sequenzen sind möglicherweise auf alternative Problemlösestrategien zurückzuführen, die entweder vom Tutor nicht unterstützt (Eingabe von Zwischenergebnissen im StoichTutor), oder aber auf Grund von komplexen Formeleingaben im ORCCA Tutor nicht erkannt werden. Analysen qualitativer Daten aus *Think Aloud* Protokollen können die Interpretation von Fehlern bereichern. In weiteren Studien werden Prozessinformationen über den Lösevorgang und die Unterstützung durch Feedback-Formen analysiert. Diese Analysen sind notwendig, um Lernunterschiede erklären zu können und um zu verstehen, von welchen Feedback-Formen der individuelle Lernende profitieren kann. Perspektivistisch sollen die ITS so verbessert werden, dass multiple Problemlösestrategien akzeptiert und Lernende durch adaptives Feedback unterstützt werden.

**Literatur**


Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. Handbook of research on learning and instruction, 2, 522-560.

Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006, June). Human Classification of Low-Fidelity Replays of Student Actions. Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, 29-36.

Bortz, J., & Döring, N. (2006). Quantitative Methoden der Datenerhebung. Forschungsmethoden und Evaluation: für Human-und Sozialwissenschaftler.

Chughtai, R., Zhang, S., & Craig, S. D. (2015, September). Usability evaluation of intelligent tutoring system: ITS from a usability perspective. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59 (1), 367-371.

Flowers, P. & Theopold, K. (2019). [ETextbook] Chemistry-2e. https://openstax.org/details/books/chemistry-2e.

Gulacar, O., Overton, T. L., Bowman, C. R., & Fynewever, H. (2013). A novel code system for revealing sources of students' difficulties with stoichiometry. Chemistry Education Research and Practice, 14 (4), 507-515.

King, E. C., Benson, M., Raysor, S., Holme, T. A., Sewall, J., Koedinger, K. R., ... & Yaron, D. J. (2022). The Open-Response Chemistry Cognitive Assistance Tutor System: Development and Implementation. Journal of Chemical Education 2022, 99, 546-552.

Marais, F., & Combrinck, S. (2009). An approach to dealing with the difficulties undergraduate chemistry students experience with stoichiometry. South African Journal of Chemistry, 62 (1), 88-96.

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011). Polite web-based intelligent tutors: Can they improve learning in classrooms?. Computers & Education, 56 (3), 574-584.

Paquette, L., de Carvahlo, A., Baker, R., & Ocumpaugh, J. (2014, July). Reengineering the Feature Distillation Process: A Case Study in Detection of Gaming the System. Educational Data mining 2014.

Schmidt, H. J. (1997). An alternate path to stoichiometric problem solving. Research in Science Education, 27, 237-249.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. Review of educational research, 85 (4), 475-511.

Wang, Z., Gong, S. Y., Xu, S., & Hu, X. E. (2019). Elaborated feedback and learning: Examining cognitive and motivational influences. Computers & Education, 136, 130-140.