Paul P. Martin[1]
David Kranz[1]
Peter Wulff[2]
Nicole Graulich[1]

[1]Justus-Liebig-Universität Gießen
[2]Pädagogische Hochschule Heidelberg

Tiefgreifende Analyse von Argumenten in der Organischen Chemie mit maschinellem Lernen

Ausgangslage

Das Bilden evidenzbasierter Argumente ist für die Entwicklung von Kommunikations- und Bewertungskompetenzen essenziell (Toulmin, 2003). In der Organischen Chemie sollten Studierende beispielsweise die Plausibilität verschiedener Reaktionswege beurteilen können, was jedoch zu Herausforderungen führt (Lieber & Graulich, 2020, 2022). Zu diesen Herausforderungen zählen das kohärente Strukturieren von Argumenten sowie die Integration chemischer Konzepte in Begründungen (Lieber, Ibraj, Caspari-Gnann & Graulich, 2022a). Um die Argumentationskompetenzen von Studierenden folglich longitudinal zu erfassen und zu fördern, bedarf es formativer Lernstandserhebungen, die mit offenen Aufgabenformaten das Bilden evidenzbasierter Argumente anleiten. Die manuelle Auswertung all dieser offenen Aufgaben ist jedoch nicht nur ressourcenintensiv, sondern auch konzeptionell schwierig, weswegen sich Methoden des maschinellen Lernens zur automatisierten Auswertung anbieten. Maschinelles Lernen ist ein Teilbereich der künstlichen Intelligenz, der Computersysteme ohne explizite Programmierung dazu befähigt, aus Daten zu lernen, Muster zu erkennen und Vorhersagen oder Entscheidungen zu treffen (Mitchell, 1997). Neben der automatisierten Auswertung von Freitext-Antworten bieten Methoden des maschinellen Lernens datengetriebene Einblicke in die Argumentationskompetenzen von Studierenden, was wiederum eine erweiterte Diagnose ermöglicht (vgl., Martin & Graulich, 2023; Zhai, Yin, Pellegrino, Haudek & Shi, 2020).

Studiendesign

Um die Argumentationskompetenzen von Studierenden der Organischen Chemie zu fördern, entwickelten Lieber, Ibraj, Caspari-Gnann und Graulich (2022a, 2022b) eine adaptive Lernumgebung, in der Studierende die Plausibilität alternativer Reaktionsprodukte beurteilen mussten. Alternative Reaktionsprodukte sind in der Organischen Chemie aufgrund miteinander konkurrierender Reaktionswege möglich, was schließlich zu mehr oder weniger plausiblen Reaktionsprodukten führt. Die Argumentation über alternative Reaktionsprodukte erfordert die Integration verschiedener chemischer Konzepte, die gegeneinander abgewogen werden müssen, um evidenzbasierte Argumente sowie Gegenargumente aufzubauen (Lieber & Graulich, 2022; Lieber, Ibraj, Caspari-Gnann, Graulich, 2022a; Watts, Park, Petterson & Shultz, 2022). Folglich bietet dieser Ansatz das Potenzial, Argumentationskompetenzen langfristig zu fördern. In der dazu von Lieber, Ibraj, Caspari-Gnann und Graulich (2022a, 2022b) entwickelten Lernumgebung entscheiden Studierende, ob das gezeigte Reaktionsprodukt plausibel ist, woraufhin sie diese Entscheidung mit Belegen und Begründungen rechtfertigen müssen. Mit diesem Aufgabenformat konnten Studierende ihre Argumentationskompetenzen signifikant verbessern (Lieber, Ibraj, Caspari-Gnann und Graulich, 2022a).

Methodischer Hintergrund und Forschungsfragen

Die in diesem Tagungsband-Beitrag vorgestellte Analyse von Argumenten in der Organischen Chemie baut auf das Studiendesign von Lieber, Ibraj, Caspari-Gnann und Graulich (2022a) auf und ist in der *computational grounded theory* verankert (Carlsen & Ralund, 2022; Nelson, 2020). *Computational grounded theory* greift auf den traditionellen Ansatz der *grounded theory* zurück, mit deren Hilfe komplexe Theorien induktiv aus Daten abgeleitet werden können. *Grounded theory* erfordert jedoch subjektive Annahmen in der Interpretation von Daten, was zu voreingenommenen Entscheidungen und begrenzter Anwendbarkeit für unstrukturierte Datensätze führen kann. Um diese Einschränkungen abzuschwächen, kombiniert *computational grounded theory* qualitative Forschung mit computergestützten Methoden, um große Datenmengen systematisch für die induktive Generierung von Theorien zu analysieren. Konkret schlägt *computational grounded theory* vier Schritte zur Datenanalyse vor, welche mit *Mustererkennung*, *Musteranpassung*, *Musterbestätigung* und *Musteraufklärung* bezeichnet werden können (Carlsen & Ralund, 2022; Martin, Kranz, Wulff & Graulich, 2023; Nelson, 2020; Tschisgale, Wulff & Kubsch, 2023).

Angelehnt an den methodischen Rahmen der *computational grounded theory* werden in diesem Tagungsband-Beitrag die folgenden Forschungsfragen (FF) beantwortet.

FF1: Welche literaturbasierten Eigenschaften sind in der datengetriebenen Analyse von schriftlichen Argumenten über alternative Reaktionsprodukte zu finden?

FF2: Mit welcher Reliabilität und Validität können die gefundenen literaturbasierten Eigenschaften automatisiert ausgewertet werden?

Eine detaillierte Beantwortung dieser Forschungsfragen ist in Martin, Kranz, Wulff & Graulich (2023) zu finden.

Ergebnisse zu FF1: Mustererkennung und Musteranpassung

Mithilfe des Clustering-Verfahrens *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) (McInnes, Healy & Astels, 2017) wurden die insgesamt 1108 Argumente, die von 64 Studierenden der Organischen Chemie II verfasst wurden, klassifiziert. Dabei konnten insgesamt 22 unterschiedliche Cluster extrahiert werden, welche im Anschluss qualitativ analysiert wurden. Innerhalb der Cluster argumentierten die Studierenden über vier unterschiedliche Themen: *Ionen und Substrate*, *Nukleophile und Elektrophile*, *Säuren und Basen* sowie *Thermodynamik und Kinetik*. Die Identifikation dieser Themen zeigt, dass die Studierenden unterschiedliche chemische Konzepte in ihre Argumentation einbetten konnten.

HDBSCAN erkannte über alle Argumente hinweg jedoch 22 unterschiedliche Cluster, weswegen dieses Verfahren die Argumente nach weiteren Merkmalen differenzierte, die während der ersten qualitativen Analyse noch nicht ersichtlich waren. Neben den Themen der Argumentation spiegelten die Cluster nämlich auch den Grad der Kausalität der angewandten chemischen Konzepte von *deskriptiv* über *relational* bis hin zu *kausal* wider (Sevian & Talanquer, 2014). Zudem verdeutlichten die Cluster auch die Tiefe der Argumentation, welche mit *phänomenologisch*, *elektronisch*, *strukturell* und *energetisch* beschrieben werden konnte (Bodé, Deng & Flynn, 2019; Deng & Flynn, 2021).

Insgesamt wird deutlich, dass mithilfe von HDBSCAN nicht nur die Themen, sondern auch der Grad der Kausalität und Tiefe der schriftlichen Argumente abgebildet werden konnte. Folglich umfassten die Ergebnisse der Cluster-Analyse die Themen *und* Komplexität der Argumentation. Auf Basis dieser Erkenntnisse konnte ein geeigneter theoretischer Rahmen für die Analyse der Argumente ausgewählt und erweitert werden. Eine Anwendung dieses

theoretischen Rahmens für die Datenauswertung zeigte schließlich, dass Studierende auf unterschiedliche chemische Konzepte und Argumentationsmuster zurückgreifen, um über die Plausibilität derselben Reaktionsprodukte zu argumentieren. Dabei konstruierten die Studierenden überwiegend *deskriptive* oder *relationale* Argumente. Während ein *struktureller* Fokus des Weiteren häufiger gesetzt wurde als ein *energetischer* Fokus, hing die Einbeziehung *elektronischer* Eigenschaften vom gegebenen Reaktionskontext ab.

Ergebnisse zu FF2: Musterbestätigung und Musteraufklärung

Zur automatisierten Auswertung der gefundenen Argumentationsmuster wurde der Datensatz im Verhältnis 65/15/20 in einen Trainings-, Validierungs- und Testdatensatz aufgeteilt. Der Trainingsdatensatz diente zum Training eines künstlichen neuronalen Netzwerkes – einer fortgeschrittenen Technik des maschinellen Lernens –, der Validierungsdatensatz diente der Kalibrierung dieses Netzwerks und der Testdatensatz diente der Bewertung der Modellgüte. Zur Analyse der schriftlichen Argumente wurde das große Sprachmodell *BERT-large-uncased* verwendet. Insgesamt konnten die Argumente mit dem so entwickelten algorithmischen Entscheidungssystem über alle 20 finalen Argumentationsmuster hinweg mit einer Reliabilität von 87% und einem Cohen's κ von 0.86 klassifiziert werden.

Um darüber hinaus das algorithmische Entscheidungssystem zu validieren, wurde mithilfe von *SHapley Additive exPlanations* (SHAP) (Lundberg & Lee, 2017) überprüft, ob die vom algorithmischen System gefundenen Entscheidungsregeln mit menschlichen Bewertungsansätzen übereinstimmen. In zahlreichen Kategorien konnte festgestellt werden, dass die Worte, die das algorithmische System zur Entscheidungsfindung nutzt, mit den im Vorhinein festgelegten menschlichen Bewertungskriterien übereinstimmen. In Kategorien, die nur mit geringerer Reliabilität ausgewertet werden konnten, lag eine Übereinstimmung zwischen menschlichen und maschinellen Bewertungsansätzen jedoch nicht vollends vor. Um diese beobachteten Diskrepanzen zu reduzieren und die Validität des algorithmischen Entscheidungssystems zu erhöhen, wurden weitere Trainingsdaten gesammelt.

Fazit, Implikationen und Ausblick

Unter Rückgriff auf den methodischen Rahmen der *computational grounded theory* konnten Methoden des maschinellen Lernens genutzt werden, um eine tiefgreifende Analyse der Argumentation von Studierenden in der Organischen Chemie zu erreichen. Die Ergebnisse der durchgeführten Studie haben gezeigt, dass…

- …die Anwendung eines Clustering-Verfahrens wie HDBSCAN dazu geeignet ist, die Themen *und* Komplexität der Argumentation von Studierenden zu bewerten.
- …die datengetriebenen Cluster bereits zahlreiche literaturbasierte Dimensionen widerspiegeln, sodass eine Klassifizierung in ein Cluster ein guter Vorhersagefaktor für eine bestimmte literaturbasierte Kategorie ist.
- …künstliche neuronale Netzwerke zur Automatisierung von komplexen, holistischen Bewertungsschemata geeignet sind.
- …Methoden des White-Boxing von Algorithmen wie SHAP genutzt werden können, um neben der Reliabilität von Algorithmen auch deren Validität aufzuklären.

Das so entwickelte algorithmische Entscheidungssystem kann zukünftig genutzt werden, um adaptives Lernen longitudinal über ein Semester in der Lehre der Organischen Chemie zu realisieren. Dabei können Studierende zu verschiedenen Zeitpunkten die Plausibilität alternativer Reaktionsprodukte beurteilen, woraufhin eine automatisierte Auswertung der Freitext-Antworten und eine Zuweisung individueller Lernangebote erfolgen kann.

Literatur

Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. Big Data & Society, 9 (1), 20539517221080146

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. Chemistry Education Research Practice, 22 (3), 749-771

Lieber, L. S., & Graulich, N. (2020). Thinking in Alternatives—A Task Design for Challenging Students' Problem-Solving Approaches in Organic Chemistry. Journal of Chemical Education, 97 (10), 3731-3738

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. Chemistry Education Research and Practice, 23 (1), 38-53

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. Chemistry Education Research and Practice, 23 (4), 811-828

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' Individual Needs Matter: A Training to Adaptively Address Students' Argumentation Skills in Organic Chemistry. Journal of Chemical Education, 99 (7), 2754-2761

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems 30. Long Beach: Curran Associates, Inc., 4765–4774

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning. Chemistry Education Research and Practice, 24 (2), 407-427

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. Journal of Research in Science Teaching. Early view article. https://doi.org/10.1002/tea.21903

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. Journal of Open Source Software, 2 (11), 205–206

Mitchell, T. M. (1997). Machine learning. New York: McGraw Hill

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. Sociological Methods & Research, 49 (1), 3–42

Toulmin, S. E. (2003). The uses of argument (Rev. Ed.). Cambridge: Cambridge University Press

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. Physical Review Physics Education Research, 19 (2), 020123-1-020123-24

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. Chemistry Education Research and Practice, 23 (2), 486–507

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. Studies in Science Education, 56 (1), 111-151