Freitextfragen digital. Automatische Bewertung und Bewerterunterstützung im Bereich der *educational applications*

Ulrike Padó

Bei Tests und Prüfungen auf Papier sind Freitextfragen ein vielseitiger und allen Beteiligten vertrauter Aufgabentyp. Sie können aber genauso bei Tests am Computer eingesetzt werden. Liegen die Antworten einmal digital vor, ist aus computerlinguistischer Sicht auch eine automatische Bewerterunterstützung möglich. Dieser Beitrag soll den Stand der Forschung beleuchten, diskutiert aber auch Einschränkungen für die Nutzung entsprechender Programme. Ziel soll sein, die Möglichkeiten der Künstlichen Intelligenz auf diesem Gebiet transparent zu machen. Denn die Nutzung im Alltag lässt noch auf sich warten: Über die Digitalisierung des Prüfens hinaus ist uns aktuell keine frei verfügbare Endanwender-Software zur Bewertungsunterstützung bekannt.

Zunächst wird der Typus »Freitextfrage« genauer definiert und seine Eigenschaften werden empirisch gegenüber anderen Fragetypen abgegrenzt. Danach geht es um die Digitalisierbarkeit des Fragetyps und die damit verbundenen Vor- und Nachteile. Schließlich soll dargestellt werden, welche Ansätze zur voll- oder teilautomatischen Bewertung von Freitextfragen im computerlinguistischen Forschungsfeld der *educational applications* existieren (dieses Feld beschäftigt sich mit der Anwendung von maschineller Sprachverarbeitung für das Lehren und Lernen).

Ein offensichtlicher Einwand gegen die Nutzung automatisierter Systeme ist die zu erwartende Fehlerquote. Zunächst wird daher festgestellt, welche Fehlerquote bei rein manueller Bewertung zu erwarten ist und damit von automatischen Systemen erreicht oder unterboten werden sollte. Dann folgt die Beschreibung eines vollautomatischen Korrektursystems. Hier ist der potentielle Verlust der menschlichen Bewertungshoheit ein wichtiges Bedenken, daher folgen zwei Vorschläge aus der Literatur zur teilautomatischen Bewerterunterstützung, bei der der Mensch das letzte Wort hat. Ein Fazit schließt den Beitrag ab.

1. Freitextfragen mit Kurzantwort

Freitextfragen sind sehr flexibel, was die erwartete Antwortlänge und die Komplexität der zu erfüllenden Aufgabe angeht. Es kann die reine Faktenreproduktion oder das Leseverstehen abgeprüft oder eine tiefgehende Analyse oder Er-

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426

örterung erfragt werden; ebenso schwanken die Antwortlängen zwischen wenigen Wörtern bis hin zu Texten von mehreren Seiten. Dieser Beitrag behandelt gezielt Freitextfragen mit Kurzantworten, d. h. mit Antworten von einer Länge von wenigen Wörtern bis ca. fünf Sätzen (Burrows et al. 2015, 61). Dies beschränkt die Komplexität der von den Prüflingen zu leistenden Aufgaben, gleichzeitig aber auch den (menschlichen und maschinellen) Korrekturaufwand. Längere und damit potentiell komplexere Antworten werden im englischsprachigen Forschungsgebiet als »Essay«-Fragen bezeichnet. Tabelle 1 zeigt zwei Beispielfragen aus verschiedenen Forschungsdatensammlungen: Die erste Zeile zeigt eine Leseverstehensfrage aus dem Fremdsprachenunterricht mit korrekter Referenzantwort und einer (leicht fehlerhaften) Prüflingsantwort mit Bewertung. Die zweite Zeile zeigt ein entsprechendes Beispiel aus der Inhaltsvermittlung zum Thema »Programmieren in Java«.

Auch wenn die Antwortlänge beschränkt wird, decken Freitextfragen Aufgaben auf verschiedenen Schwierigkeitsstufen ab. Für die Abschätzung, wie schwierig eine Prüfungsfrage für die Prüflinge ist, wird gern die sog. Bloom'sche Taxonomie (Anderson et al. 2014) hinzugezogen. Sie definiert eine Dimension kognitiver Prozesse, die beim Beantworten der Frage gefordert sein können, und ordnet sie der Schwierigkeit nach. Der am wenigsten aufwändige kognitive Prozess nach Anderson et al. (2014) ist »Remember«, die reine Reproduktion von Wissen. Es folgt »Understand« (Konzepte erklären und in Zusammenhang setzen können) und »Apply« (Anwenden im neuen Kontext), gefolgt von den anspruchsvolleren Stufen »Analyze« (Differenzieren, Strukturieren, Ableiten), »Evaluate« (auf Konsistenz, Anwendbarkeit und Angemessenheit prüfen) und »Create« (selbständig Neues schaffen).

Für die Bewertung von Leseverständnisfragen schlagen Day und Park (2005) eine eigene Taxonomie vor: Das Auffinden von Information aus dem Text ist auf Stufe »Literal«, das Kombinieren von Informationen aus verschiedenen Stellen im Text ist auf Stufe »Reorganization«. Werden für die Antwort Inferenzen aus Textinformationen gezogen, entspricht dies der Stufe »Inference«. »Prediction« bezeichnet die Aufgabe, aus dem Text und dem außertextuellen Wissen über die Welt die weitere Entwicklung der Lesegeschichte vorherzusagen; »Evaluation« bezeichnet die Einordnung des Textes in das eigene bisherige Wissen und »Personal Response« fragt nach einer Reflektion der eigenen emotionalen Reaktion auf den Text.

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426

Tabelle 1: Beispielfragen und -antworten (letztere ohne Fehlerkorrektur wiedergegeben) mit Fokus auf Sprachvermittlung (CREG) und Inhaltsvermittlung (CSSAG)

Quelle	Frage	Referenzantwort	Prüfling	Bewertung
CREG	Wie kann man sich in Pillnitz erholen und den Stress vergessen?	Bei einem Bummel durch den Park kann man den Stress vergessen. ODER Man kann die Schönheit genießen.	»Man kann eine Bummel durch den weitläufigen Park machen.«	korrekt
CSSAG	Wenn eine Methode einen Vertrag zwischen Programmierer und Benutzer darstellt, welche Fehlerarten gibt es dann?	1. Programmierfehler: Die Methode gibt unerwartete Ergebnisse zurück. 2. Benutzerfehler: Der Benutzer verwendet ungeeignete Argumente.	»Der Benutzer hat die Klasse falsch benutzt. Der Programmierer hat den Code falsch codiert.«	korrekt

Freitextfragen mit Kurzantwort decken aufgrund der Längenbeschränkung der Antworten diese Taxonomien empirisch jeweils bis zur Mitte ab. Dies spiegelt sich in zwei existierenden Datensammlungen für das Deutsche, CREG (»Corpus of Reading Comprehension Exercises in German«, vgl. Meurers et al. 2011, Leseverständnisfragen für den DaF-Unterricht) und CSSAG (»Computer Science Short Answers for German«, vgl. Kiefer/Padó 2015, Inhaltsfragen zum Thema »Programmierung«). In CREG sind 79 % der 1032 Fragen auf der Stufe »Literal«, 13 % auf der Stufe »Reorganization« und 8 % auf der Stufe »Inference«. Die 31 Wissensfragen in CSSAG sind laut Padó (2017) zu 52 % auf der Stufe »Remember«, zu 38 % auf der Stufe »Understand« und zu 10 % auf der Stufe »Apply«[1]. Auf den höheren Day&Park- bzw. Bloom-Stufen gibt es jeweils keine Fragen. Diese Stufen sind selbstverständlich auch mit Freitextfragen abzudecken, allerdings sind deutlich längere Antworten zu erwarten. Solche Fragen fallen also in den »Essay«-Bereich.

Die Beschränkung auf die unteren bis mittleren Taxonomiestufen hat Auswirkungen für die digitale Umsetzung von Freitextfragen: Auf den unteren Taxonomiestufen nach Bloom machen Multiple-Choice-Fragen den Freitextfragen je nach Fachgebiet und Prüfungskultur Konkurrenz. McMillan (2001) er-

1 Bei der Zuordnung der Fragen zu den Stufen ist es wichtig, nicht nur auf die von Anderson et al. (2014) vorgeschlagenen Formulierungen zu achten, sondern auch darauf, ob die gewünschte Antwort bereits im vorausgehenden Unterricht explizit eingeführt wurde: Dadurch reduziert sich die reale Taxonomiestufe der Frage auf »Remember«.

mittelt für Lehrende an weiterführenden Schulen in den Jahrgangsstufen 6–12 im US-Bundesstaat Virginia, dass dort für Aufgaben auf der Bloom-Stufe »Remember« überwiegend Multiple-Choice-Fragen verwendet werden und Freitextantworten vermehrt in den höheren Jahrgangsstufen sowie öfter in den geisteswissenschaftlichen als den naturwissenschaftlichen Fächern eingesetzt werden. Gerade im Kontext digitaler Prüfungen ist dies nachvollziehbar, denn Multiple-Choice-Aufgaben lassen sich einfach automatisiert korrigieren. Gleichzeitig erlauben Freitextfragen den Prüfenden Einblick in die Gedanken und möglichen Irrtümer der Prüflinge, was bei Multiple-Choice-Aufgaben nicht der Fall ist. Für Prüfungen, die nicht nur summativ bewerten, sondern auch eine Rückmeldung zur Leistungsverbesserung liefern sollen, sind Freitextfragen daher trotz ihres weit höheren Korrekturaufwands attraktiv. Gleiches gilt, wenn die schriftliche Ausdrucksfähigkeit der Prüflinge bewertet werden soll.

2. Digitale Freitextfragen

Technisch gesehen können Freitextfragen problemlos in digitalen Prüfungen verwendet werden. Die an vielen Schulen und Hochschulen eingesetzten Learning-Management-Systeme (LMS) wie z. B. Moodle[2] oder ILIAS[3] bieten Module für die Durchführung von Online-Tests an und stellen auch den Fragetyp »Freitextfrage« zur Verfügung. Somit lässt sich in der Praxis überall dort niederschwellig digital prüfen, wo ein LMS bereits im Einsatz ist. Aufgrund von prüfungsrechtlichen Anforderungen an Archivierbarkeit, Systemstabilität und Einbruchssicherheit für summative Prüfungen sind Tests im LMS ohne flankierende Maßnahmen allerdings am besten für formatives Prüfen geeignet.

Dabei verspricht die reine Digitalisierung von Freitextfragen bereits eine Beschleunigung der Bewertung. Schulz und Apostolopoulos (2011) nennen eine Verkürzung der Korrekturzeit von digital gestellten Freitextfragen um 33 % im Vergleich zur Durchführung auf Papier. Sie führen dies hauptsächlich darauf zurück, dass die getippten Antworten im Gegensatz zu handschriftlichem Text immer mühelos lesbar sind.

Allerdings setzt die digitale Umsetzung voraus, dass alle Prüflinge zügig tippen können. Sonst entsteht durch den längeren Zeitbedarf beim Schreiben für langsamere Tipper faktisch eine Ungleichbehandlung. Diese kann durch Verlängerung der Bearbeitungszeit ausgeglichen oder dadurch vermieden werden, dass die Prüflinge ausdrücklich auf die digitale Prüfung vorbereitet werden.

Die Korrektur digital vorliegender Antworten verspricht auch Vorteile für die Objektivität der menschlichen Bewertung: Zum einen besteht die Möglichkeit der Anonymisierung, da die Bewertenden die Prüflinge nicht an der Handschrift

2 www.moodle.org.
3 https://www.ilias.de/.

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426

erkennen. Zum anderen ist es einfach möglich, alle Antworten auf dieselbe Frage nacheinander zu korrigieren, so dass sie direkt miteinander verglichen werden können und so die Bewertung konsistenter wird.

Die Bewertungszeit lässt sich durch einfache Maßnahmen weiter verkürzen: Die Optimierung der Reihenfolge, in der die Antworten auf eine Freitextfrage präsentiert werden, beschleunigt ebenfalls die Korrektur. Dies gilt insbesondere für solche Fragen, deren Antworten sonst besonders langwierig zu korrigieren sind (Padó/Kiefer 2015). Für die Sortierung werden die Antworten mit einer hinterlegten Musterlösung verglichen und anhand ihrer Ähnlichkeit mit dieser sortiert. Die Ähnlichkeit zur Musterlösung wird mit Hilfe informatischer Vergleichsalgorithmen für Buchstabenketten ermittelt. So können die ähnlichsten (und damit wahrscheinlich richtigen) Antworten gemeinsam präsentiert werden; auch die der Musterlösung unähnlichsten Antworten stehen zusammen. Dies sind oft leere Antworten oder Kommentare wie »keine Ahnung«. Dazwischen finden sich Zweifelsfälle, die zwar Übereinstimmungen mit der Musterlösung aufweisen, aber ihr nicht vollständig entsprechen.

Die bisher genannten Maßnahmen verändern das grundsätzliche Vorgehen bei der Bewertung von Freitextfragen nicht: Alle Antworten werden von menschlichen Korrektoren gelesen und bewertet. Eine Effizienzsteigerung ergibt sich zunächst nur durch die bessere Lesbarkeit oder eine optimierte Präsentationsreihenfolge.

3. Menschliche Bewertung als Maßstab

Wenn Prüfungsantworten einmal digital vorliegen, ist auch eine vollautomatische Bewertung denkbar. Menschlicher Korrekturaufwand würde durch den Einsatz eines maschinellen Bewertungssystems wegfallen, Freitextfragen wären dann ähnlich unaufwendig zu korrigieren, wie es Multiple-Choice-Fragen jetzt schon sind.

Auch manche Inkonsistenzen in der menschlichen Bewertung können durch ein automatisches System vermieden werden: Sie entstehen durch Ermüdung oder Beurteilungsfehler wie den Hofeffekt eines herausstechenden Merkmals oder den Primacy-Fehler, bei dem ein früher Eindruck die spätere Bewertung zu stark bestimmt (vgl. z. B. Schwaighofer et al. 2019). Mieskes und Padó (2018) überprüften die gängigen Forschungsdatensammlungen für Freitextfragen auf die Qualität der zugehörigen menschlichen Bewertung hin. Datensammlungen, die Fragen und Antworten aus formativen Tests während der Kurslaufzeit mit ihrer vom Lehrenden gegebenen Bewertung sammeln, zeigten eine Korrekturpräzision von ca. 85 % – das heißt, 15 % der Bewertungen waren falsch (gemessen an der hinterlegten korrekten Bewertung). Ein Datensatz, der Antworten aus besonders sorgfältig korrigierten summativen Tests enthält, zeigte eine weit höhere Korrekturpräzision von 94 %, enthält so aber immer noch 6 % falsche (oder zumin-

dest diskussionswürdige) Bewertungen. Dabei ist auch zu bedenken, dass die Korrekturpräzision sinkt, wenn nicht nur als »richtig« und »falsch« bewertet wird, sondern mehrere Notenstufen zu berücksichtigen sind. Je mehr Bewertungsmöglichkeiten vorliegen, desto mehr Gelegenheiten für Fehler oder diskussionswürdige Entscheidungen gibt es bekanntlich.

Diese Ergebnisse für die menschliche Bewertung geben demnach einen Korridor vor, in dem sich die Korrekturpräzision eines vollautomatischen Systems mindestens bewegen sollte, um keinen Qualitätsverlust zu verursachen. Fehlerfrei sind automatische Bewertungssysteme dennoch ebenso wenig wie der Mensch. Dies ist ein zulässiger grundsätzlicher Einwand gegen ihre Verwendung, muss aber durchaus differenziert gesehen werden, wie der nächste Abschnitt erläutert.

4. Ansätze zur automatischen Bewertung

Für die Erstellung automatischer Bewertungen werden Methoden der Künstlichen Intelligenz, nämlich maschinelle Lernsysteme, eingesetzt. Wir betrachten nun die Funktionsweise dieser Systeme und den Trainingsprozess genauer: Aus technischer Sicht wird die Bewertung von Freitextfragen meist als »Klassifikation« betrachtet, Ziel ist also die Zuweisung einer Bewertung aus einer Liste von möglichen Bewertungen. Dies können Notenstufen sein oder nur die Bewertungen »bestanden« und »nicht bestanden«. Das Training des Lerners erfolgt ›überwacht‹, also aufgrund von Beispieldaten, für die die korrekte Zielklasse bekannt ist.

Die Daten werden vorverarbeitet, um relevante Eigenschaften (engl. *features*) der Fragen und Antworten für das Lernsystem sichtbar zu machen.[4] Features können von verschiedenen computerlinguistischen Verarbeitungsebenen stammen, wie der Zeichenebene, der Wortebene, der Syntax und auch einer Annäherung an die Semantik der Antworten.

Obwohl die Systeme die Antworten nicht wie ein Mensch verstehen, berücksichtigen sie also dennoch Wörter und ihre syntaktischen und semantischen Beziehungen zueinander. So kann zum Beispiel das Vorkommen oder Fehlen von Negation in einer Aussage erkannt werden.

Sind die Features der Beispieldaten berechnet, stehen verschiedene mathematische Algorithmen zur Verfügung, um die Klassifikationsfunktion zu ermitteln. Diese leistet die eigentliche Prädiktion: Sie bekommt die Features einer Antwort als Eingabe und liefert eine Bewertungsvorhersage als Ausgabe. Der gewählte Algorithmus bestimmt, wie die Klassifikationsfunktion genau definiert

4 Deep-Learning-Verfahren (neuronale Netze) ersetzen das Berechnen von Features durch eine komplexe Systemarchitektur, deren Training viele Beispieldaten erfordert. Aufgrund der relativ geringen Datenmengen ist Deep Learning für die Freitextfragenbewertung im Moment nicht relevant (Riordan et al. 2017).

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426

wird. Überwachten Klassifikationsverfahren ist gemein, dass sie die Funktion anhand von Trainingsdaten, deren korrekte Bewertung bekannt ist, spezifizieren. Hierfür werden die Beispieldaten in einen größeren Trainingsteil und einen kleineren Testteil aufgeteilt. Der Trainingsteil dient zur Induktion der Klassifikationsfunktion nach Maßgabe des gewählten Algorithmus, indem zum Beispiel Parameter erhoben oder optimiert werden. Die voraussichtliche Korrekturpräzision der Funktion wird anhand des Testteils bestimmt: Die Features der im Training nicht verwendeten Testantworten dienen für die Vorhersage von Bewertungen. Ob diese korrekt sind, lässt sich ja anhand der gegebenen Zielbewertung prüfen.

Die Auswahl des konkreten Algorithmus erfolgt entweder anhand eines Ergebnisvergleichs verschiedener Algorithmen oder aufgrund bestimmter Eigenschaften des Algorithmus, zum Beispiel wie einfach es ist, die Vorhersagen des Algorithmus nachzuvollziehen.

Mit diesem Vorgehen erreichen maschinelle Lernsysteme schon länger durchaus Korrekturpräzisionen, die den oben beschriebenen 15 % Fehleranteil bei der menschlichen Korrektur entsprechen (Dzikovska et al. 2013). Allerdings verbirgt die Kennzahl der Korrekturpräzision möglicherweise einen *algorithmic bias,* also einen Beurteilungsfehler der Maschine, der in ihrer Funktionsweise begründet liegt. Ein automatisches Bewertungssystem muss also auch daraufhin überprüft werden, dass es nicht durch solch einen Fehler bestimmte Lernergruppen bevorzugt oder benachteiligt.

Algorithmic bias entsteht im maschinellen Lernen gerade durch dasjenige Abstützen auf empirisch erhobenen Trainingsdaten, durch das Objektivität garantiert werden soll. Zum einen sind oftmals ungleich viele Beispiele für die verschiedenen Zielbewertungen vorhanden. So gibt es für einfache Freitextaufgaben natürlich viel mehr richtige als falsche Antworten, für schwere Aufgaben mehr falsche als richtige Antworten. Maschinelle Lerner funktionieren aber je besser, desto mehr Trainingsbeispiele sie bekommen. Es ist daher zu beobachten, dass die häufigste Zielbewertung aufgrund der guten Datenbasis am verlässlichsten gegeben wird und seltene Zielbewertungen weniger verlässlich sind, weil zu wenig Daten im Training vorlagen (vgl. z. B. Mieskes/Padó 2018). Durch die sorgfältige Auswahl der Trainingsdaten lässt sich diese Verzerrung (wohl nur) teilweise ausgleichen (Loukina et al. 2019).

Eine zweite Quelle von *algorithmic bias* ist das mögliche Vorhandensein von Verzerrungen in der menschlichen Bewertung der zugrundeliegenden Trainingsdaten. Hierbei geht es wieder um die Häufigkeit, mit der eine Zielbewertung in den Trainingsdaten vorkommt, jetzt aber heruntergebrochen auf bestimmte Benutzergruppen. Werden zum Beispiel alle Antworten mit einer bestimmten Eigenschaft – z. B. dem Vorkommen eines bestimmten Vokabulars – unabhängig von ihrer objektiven Korrektheit von den menschlichen Bewertern schlechter bewertet, werden Antworten mit dieser Eigenschaft auch von der Maschine schlechter bewertet werden – weil es die Trainingsdaten so vorgeben. Diese Art

von Maschinenbias entspricht also direkt einer Bewertungsverzerrung durch Menschen, die ja eigentlich durch den Einsatz von maschinellen Verfahren vermieden werden soll.

Wenn maschinelle Lernverfahren eingesetzt werden sollen, ist also eine sorgfältige Prüfung ihres Bewertungsverhaltens unabdingbar. Dies geschieht am besten, indem die zukünftige Nutzerin die Vorhersagen der Maschine für einen eigenen Datensatz mit dafür vorliegenden menschlichen Bewertungen abgleicht. Wichtig ist nicht nur, wie hoch die Abweichungen insgesamt sind, sondern auch, ob sie eine bestimmte Tendenz zeigen (z. B. eine allgemein zu positive oder zu negative Maschinenbewertung) oder ob sonst in den Abweichungen Muster auffallen. Im nächsten Schritt kann die Nutzerin dann abwägen, ob die zu erwartende Korrekturpräzision für das gewünschte Einsatzszenario angemessen ist. So kann für einen Selbsttest im Rahmen von Blended Learning eine gewisse Fehlerquote akzeptabel sein, wenn dadurch den Lernenden ermöglicht wird, jederzeit zu arbeiten und nicht auf Rückmeldung der Lehrenden warten zu müssen. Für eine summative Prüfung mit großer Bedeutung für die Prüflinge muss gegebenenfalls schon aus prüfungsrechtlichen Gründen die (mehrfache) menschliche Bewertung verwendet werden.

Eine intensive Ergebnisanalyse mildert auch einen weiteren wichtigen Einwand gegen die Verwendung automatisierter Verfahren ab, nämlich die Sorge, damit die Bewertungshoheit aus der Hand zu geben und intransparent erstellte Bewertungen akzeptieren zu müssen.

Eine andere effektive Maßnahme zum Einsparen von menschlichem Korrekturaufwand ohne die völlige Aufgabe menschlicher Kontrolle über die Bewertung ist die Nutzung automatischer Verfahren zur Bewerterunterstützung. Diese Verfahren nehmen eine Vorgruppierung der Antworten vor oder schlagen vorläufige Bewertungen vor, die endgültige Bewertungsentscheidung liegt aber weiterhin beim Menschen.

5. Ansätze zur automatischen Bewerterunterstützung

In der Forschungsliteratur gibt es aktuell zwei grundsätzliche Herangehensweisen zur Bewertungsunterstützung. Das erste Vorgehen kommt völlig ohne automatische Bewertungssysteme aus: Brooks et al. (2014) und Horbach et al. (2014) schlagen vor, die Kurzantworten nach ihrer Ähnlichkeit zueinander in Gruppen (Cluster) zu sortieren. Die Cluster werden hierbei nicht automatisch bewertet. Menschliche Bewerter können vielmehr pro Cluster eine Bewertung vergeben, statt jede Antwort einzeln bewerten zu müssen. Die Gruppierung erlaubt es den menschlichen Bewertenden, sich einen schnellen Überblick über die vorliegenden Antwortvarianten zu verschaffen. Dies hat für formative Prüfungen den Vorteil, dass typische Interpretationsvarianten oder Missverständnisse so schnell sichtbar

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426

und deutlich quantifiziert werden und in der nächsten Unterrichtseinheit aufgegriffen werden können.

Horbach et al. (2014) evaluieren ihr Verfahren anhand von Antworten auf Lese- und Hörverständnisfragen aus Einstufungstests im Bereich Deutsch als Fremdsprache. Kriterien für die Gruppierung sind, wie ähnlich sich die verwendeten Wörter und Formulierungen der einzelnen Antworten sind. Horbach et al. geben die Korrekturpräzision bei der Verwendung ihres Verfahrens mit mindestens 85 % an (dies entspricht der oben berichteten empirisch beobachteten Korrekturpräzision ohne Unterstützung). Gleichzeitig mussten die Bewertenden nur durchschnittlich 40 % der Antworten manuell bearbeiten. Horbach et al. beobachten allerdings je nach Frage starke Abweichungen von diesem Durchschnittswert (manchmal sind nur 20 % Antworten zu korrigieren, manchmal bis zu 80 %).

Brooks et al. (2014) berichten mit 25 Bewertenden auf englischsprachigen Daten ähnliche Ergebnisse: Die Korrekturpräzision bleibt vergleichbar zum Vorgehen ohne Unterstützung. Gleichzeitig sinkt der Bewertungsaufwand auf ca. 33 %. Der Ansatz bietet also einen hohen Grad an menschlicher Kontrolle über die Noten bei einer spürbaren Effizienzsteigerung.

Ein anderes Vorgehen, das vollautomatische Bewertungssysteme für die Unterstützung menschlicher Bewerter verwendet, schlagen Mieskes und Padó (2018) vor: Ein maschinelles *Ensemble*, also eine Kombination aus drei verschiedenen automatischen Bewertern, schlägt Bewertungen vor. Dazu sagt jeder automatische Bewerter einzeln eine Bewertung voraus. Die am häufigsten vergebene Bewertung wird vorgeschlagen. Nun können diejenigen Bewertungen identifiziert werden, die möglicherweise nicht zuverlässig sind und vom Menschen korrigiert werden sollten: Waren alle drei Bewertungsvorschläge gleich, ist die Vorhersage nachweislich ähnlich verlässlich wie die menschliche Bewertung. Sind sich nur zwei Maschinen einig oder liegen gar drei unterschiedliche Vorhersagen vor, ist die Vorhersage weniger verlässlich.

Die Entscheidung für die Übernahme der Korrektur durch den Menschen kann in verschiedenen Abstufungen gefällt werden: Sollen nur Fälle betrachtet werden, in denen die Maschinen sich völlig uneinig waren, oder auch solche, in denen es eine abweichende Vorhersage gab? Je häufiger der Mensch eingreift, desto geringer die Arbeitsersparnis, aber auch desto höher die Verlässlichkeit der finalen Bewertungen. Je nach Einsatzszenario können so menschlicher Zeitaufwand und Bewertungsgenauigkeit austariert werden: weniger Aufwand für wöchentliche Kurztests, mehr Aufwand und mehr Genauigkeit für Tests mit höherem Gewicht. Die theoretisch mögliche Aufwandsersparnis bei einer Korrekturpräzision von 85 % variiert von Datensatz zu Datensatz; Einsparungen von 70–80 % des Arbeitsaufwands sind typisch.

Leider gibt es noch für keine der beiden beschriebenen Herangehensweisen Software für Endanwender. Ein Grund ist, dass Forschungsgruppen im Allge-

meinen die Finanzmittel für die Entwicklung und den dauerhaften Support von Endanwender-Software fehlen.

6. Fazit

Dieser Beitrag beschäftigte sich mit Freitextfragen, auf die eine bis zu fünf Sätze kurze Antwort zu erwarten ist. Aufgrund der Längenbeschränkung ergibt sich, dass sich Freitextfragen mit Kurzantworten hauptsächlich für leichte bis etwa mittelschwere Aufgaben eignen; dies wurde auch empirisch an zwei Forschungsdatensammlungen verifiziert. Trotz dieser Einschränkung sind Freitextfragen in vielen Lehr-Lern-Situationen flexibel einsetzbar und können ebenso digital wie herkömmlich auf Papier gestellt werden. Ein deutlicher Nachteil gegenüber anderen digital nutzbaren Fragetypen ist aber der manuelle Korrekturaufwand. Hier verspricht schon digitalisiertes Prüfen allein eine Beschleunigung der Korrektur, da die Antworten mühelos lesbar sind. Dies setzt aber voraus, dass der Umgang mit Maus und Tastatur allen Prüflingen vertraut und möglich ist.

Liegen die Prüfungsantworten einmal digital vor, so ist auch eine computergestützte Korrektur möglich. Die Funktionsweise vollautomatischer Bewertungsmethoden wurde erklärt und ihre Fehlerquote mit der Quote bei rein manueller Bewertung in Beziehung gesetzt, die sich aus der Auswertung von Forschungsdatensätzen ergibt. Nachteil eines vollautomatischen Systems ist, dass der Mensch die Bewertungshoheit völlig aus der Hand gibt und menschliche Bewertungsfehler potentiell gegen systematische Verzerrungen der maschinellen Bewertung eingetauscht werden. Daher folgte eine Beschreibung zweier Vorschläge zur teilautomatischen Bewerterunterstützung: Ein Verfahren nutzt gar keine vollautomatischen Bewertungsvorschläge, das andere beruht auf der gezielten Identifikation und manuellen Nachkorrektur voraussichtlich falscher automatischer Bewertungen.

Dieser Beitrag hatte zum Ziel, den momentanen Stand der Forschung im Bereich der Bewertung von Freitextfragen darzustellen. Dabei wurde deutlich, dass die Nutzung eines automatischen oder teilautomatischen Systems unbestreitbaren Nutzen hat, aber sorgfältig vorbereitet werden sollte. Zunächst ist es wichtig, sich über die benötigte Korrekturpräzision und den notwendigen Grad an menschlicher Bewertungshoheit im Klaren zu sein. Dann muss ein automatisches System sorgfältig überprüft werden, damit seine Fehlerquote und mögliche Bewertungsverzerrungen in die Abwägung einbezogen werden können. Je nach Situation ist so gar kein, ein begrenzter oder auch der vollständige Einsatz einer automatischen Methode ratsam und sinnvoll.

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426

Literaturverzeichnis

Anderson, Lorin W./Krathwohl, David R./Airasian, Peter W./Cruikshank, Kathleen A./Mayer, Richard E. (Hrsg.) (2014): A Taxonomy for Learning, Teaching and Assessing. A revision of Bloom's Taxonomy of Educational Objectives. Harlow: Pearson.

Brooks, Michael/Basu, Sumit/Jacobs, Charles/Vanderwende, Lucy (2014): Divide and Correct. Using Clusters to Grade Short Answers at Scale. In: Sahami, Mehran (Hrsg.): L@S 2014. Proceedings of the first ACM conference on Learning @scale. Atlanta: Association for Computing Machinery. S. 89–98.

Burrows, Steven/Gurevych, Iryna/Stein, Benno (2015): The Eras and Trends of Automatic Short Answer Grading. In: International Journal of Artificial Intelligence in Education 25. S. 60–117.

Day, Richard R./Park, Jeong-Suk (2005): Developing Reading Comprehension Questions. In: Reading in a Foreign Language 17. S. 60–73.

Dzikovska, Myroslava O./Nielsen, Rodney D./Brew, Chris/Leacock, Claudia/Giampiccolo, Danilo/Bentivogli, Luisa/Clark, Peter/Dagan, Ido/Trang Dang, Hoa (2013): SemEval-2013 Task 7. The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In: Diab, Mona T./Baldwin, Timothy/Baroni, Marco (Hrsg.): Second Joint Conference on Lexical and Computational Semantics (*SEM). Bd. 2: Seventh International Workshop on Semantic Evaluation. Atlanta: Association for Computational Linguistics. S. 263–274.

Horbach, Andrea/Palmer, Alexis/Wolska, Magdalena (2014): Finding a Tradeoff between Accuracy and Rater's Workload in Grading Clustered Short Answers. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Loftsson, Hrafn/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hrsg.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik: European Language Resources Association. S. 588–595.

Kiefer, Cornelia/Padó, Ulrike (2015): Freitextaufgaben in Online-Tests. Bewertung und Bewertungsunterstützung. In: HMD Praxis der Wirtschaftsinformatik 52. S. 96–107.

Loukina, Anastassia/Madnani, Nitin/Zechner, Kaus (2019): The Many Dimensions of Algorithmic Fairness in Educational Applications. In: Yannakoudakis, Helen/Kochmar, Ekaterina/Leacock, Claudia/Madnani, Nitin/Pilán, Ildikó/Zesch, Torsten (Hrsg.): Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Stroudsburg: Association for Computational Linguistics. S. 1–10.

McMillan, James H. (2001): Secondary Teachers' Classroom Assessment and Grading Practices. In: Educational Measurement. Issues and Practice 20. S. 20–32.

Meurers, Detmar/Ziai, Ramon/Ott, Niels/Kopp, Janina (2011): Evaluating Answers to Reading Comprehension Questions in Context. Results for German and the Role of Information Structure. In: TextInfer 2011 Workshop on Textual Entailment. Proceedings of the Workshop. Stroudsburg: Association for Computational Linguistics. S. 1–9.

Mieskes, Margot/Padó, Ulrike (2018): Work Smart. Reducing Effort in Short-Answer Grading. In: Proceedings of the seventh workshop on NLP for computer-assisted language learning. Linköping Electronic Conference Proceedings 152. S. 57–68.

Padó, Ulrike/Kiefer, Cornelia (2015): Short Answer Grading: When sorting helps and when it doesn't. In: Volodina, Elena/Borin, Lars/Pilán, Ildikó (Hrsg.): Proceedings of the fourth workshop on NLP for computer-assisted language learning. Linköping: Linköping Electronic Conference Proceedings 114. S. 42–50.

Padó, Ulrike (2017): Question Difficulty. How to Estimate without Norming, how to Use for Automated Grading. In: Tetreault, Joel/Burstein, Jill/Leacock, Claudia/Yannakoudakis, Helen (Hrsg.): The Twelfth Workshop on Innovative Use of NLP for Building Educational Applications. Proceedings of the Workshop. Stroudsburg: Association for Computational Linguistics. S. 1–10.

Riordan, Brian/Horbach, Andrea/Cahill, Aoife/Zesch, Torsten/Lee, Chong Min (2017): Investigating neural architectures for short-answer scoring. In: Tetreault, Joel/Burstein, Jill/ Leacock, Claudia/Yannakoudakis, Helen (Hrsg.): Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. Stroudsburg: Association for Computational Linguistics. S. 159–168.

Schulz, Alexander/Apostolopoulos, Nicolas (2011): Potenziale computergestützter Prüfungen. In: Hamburger eLearning-Magazin 7. S. 37–39.

Schwaighofer, Matthias/Heene, Moritz/Bühner, Markus (2019): Grundlagen und Kriterien der Diagnostik. In: Urhane, Detlef/Dresel, Markus/Fischer, Frank (Hrsg.): Psychologie für den Lehrberuf. Berlin: Springer. S. 471–491.

Prof. Dr. Ulrike Padó, Hochschule für Technik Stuttgart, ulrike.pado@hft-stuttgart.de

Mitteilungen des Deutschen Germanistenverbandes 2/2020, Jg. 67, ISSN 0418-9426