Am Ende jedes Semesters das gleiche Bild: ein riesiger Stapel Klausuren wartet schon seit Wochen darauf benotet zu werden. Für jede Antwort ein ermutigendes und korrigierendes Feedback zu schreiben, würde noch mehr Zeit kosten. Außerdem kursiert die Klausur im Internet. Für nächstes Jahr müssen dann doch mal neue Aufgaben her.

Künstliche Intelligenz in der Bildung

Wie Sprachtechnologie Lehrende unterstützen kann

Von Torsten Zesch, Andrea Horbach & Ronja Laarmann-Quante

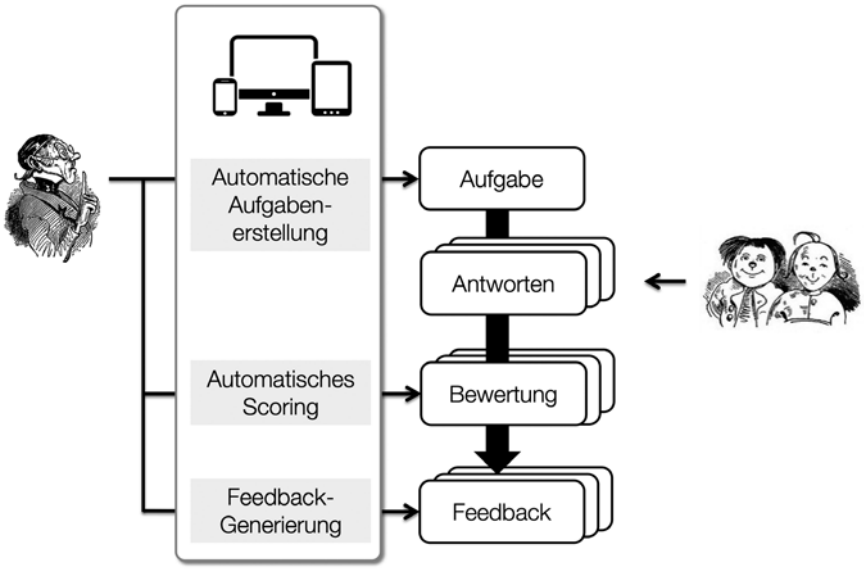
Die Darstellung oben ist bewusst überspitzt, aber wer unter den Lehrenden kennt nicht das Gefühl, zu wenig Zeit für die Lernenden zu haben. Daher ist es nicht verwunderlich, dass künstliche Intelligenz (KI) als Möglichkeit zur Unterstützung und Entlastung der Lehrenden in Betracht gezogen wird.

Abbildung (1) veranschaulicht, welche Prozesse im Lehralltag hauptsächlich ablaufen und von KI-Unterstützung profitieren können. Lehrende erstellen Aufgaben, seien es Übungs- oder Klausuraufgaben. Die Lernenden bearbeiten die Aufgaben und bekommen dazu eine Rückmeldung, entweder in Form einer Bewertung (richtig/falsch bzw. Punktzahl) oder besser in Form von ausführlichem Feedback. Natürlich ist diese Darstellung vereinfacht – Studierende werden zum Beispiel auch Fragen stellen und und mit den Lehrenden oder auch untereinander diskutieren – dennoch deckt die Darstellung die drei wichtigsten Prozesse ab: Aufgabenerstellung, Bewertung und Feedback. Künstliche Intelligenz kann nun eingesetzt werden, um die Lehrenden in allen drei Schritten zu entlasten.

Was immer Lehrende oder Lernende tun, das meiste davon wird verbal kommuniziert, einfach weil es für uns Menschen die natürlichste Art der Interaktion ist. Lehrende erklären eine Theorie in Worten, Studierende schreiben ein Essay und bekommen Feedback dazu oder beantworten Fragen in einer Klausur. Dabei bedienen wir uns verschiedener Modalitäten, gesprochener und geschriebener Sprache, die von Hand oder am Computer geschrieben wird. Zur KI-Unterstützung eignen sich hier also insbesondere Methoden der Sprachtechnologie.

Wir werden in diesem Artikel einen Überblick über unsere Forschung in diesem Gebiet geben und insbesondere darauf eingehen, wie die drei Teilbereiche – Aufgabenerstellung, Bewertung und Feedback – mit sprachtechnologischen Methoden unterstützt werden können. Zusätzlich zur technologischen Herausforderung diskutieren wir die gesellschaftlichen Implikationen, ins-

Torsten Zesch. Foto: Vladimir Unkovic



(1) Anwendungsgebiete von Sprachtechnologie in der Lehre.
Quelle: eigene Darstellung

besondere die möglichen ethischen Probleme, die der Einsatz solcher Technologien im Bildungsbereich nach sich ziehen könnte.

Aufgabenerstellung

Der erste Schritt im Lehr-Lernprozess aus Sicht der Lehrenden ist die Erstellung von Aufgaben. Idealerweise sollten diese Aufgaben einen für die Lernenden angemessenen Schwierigkeitsgrad haben, oder sogar individuell angepasst sein. Wir stellen im Folgenden verschiedene Aufgabentypen vor, für die eine automatische Erstellung und Individualisierung eine Herausforderung darstellt, die aber dennoch mit sprachtechnologischen Methoden bewältigt werden können.

Leseverständnisfragen

Leseverständnisfragen zu Lehrbuchtexten helfen Lernenden bei der vertiefenden Verarbeitung der Inhalte. Darüber hinaus kann das Verständnis von vorlesungsbegleitender Literatur mit Leseverständnisfragen geprüft werden. Auch beim Lernen einer Fremdsprache profitieren Studierende davon, Fragen zu einem Text gestellt zu bekommen und Antwortsätze aktiv formulieren zu müssen.

Die manuelle Erstellung von Leseverständnisfragen ist jedoch ein zeitaufwändiger Prozess. Hier könnten die Lehrenden unterstützt werden, indem Fragen automatisch generiert und zur Auswahl angeboten werden. Dafür starten wir mit dem Lehrbuchtext, wählen geeignete, wichtige Textstellen aus und transformieren diesen Text dann in eine Frage. Damit die Transformation automatisch durchgeführt werden kann, muss das System Wissen über die Satzstellung (Syntax) in Fragen und Aussagesätzen besitzen. So steht im Deutschen in einem Fragesatz zu einer W-Frage (wer, was, wo?) das Verb an zweiter Stelle hinter dem Fragewort.

Betrachten wir dazu den folgenden Satz, der so in einem Text für den Deutsch-als-Fremdsprache-Unterricht stehen könnte: *Marie stand heute früh um 7 Uhr auf und trank einen Kaffee mit viel Milch.* Nehmen wir weiter an, wir möchten eine Frage zu dem Kaffee stellen. Wir müssen zunächst erkennen, dass der Satz eigentlich aus zwei Einzelsätzen besteht. *Marie stand heute früh um 7 Uhr auf* und *Marie trank einen Kaffee mit viel Milch*, und dass auch das Subjekt des zweiten Satzes „*Marie*" ist, obwohl das, oberflächlich betrachtet, gar nicht im Text steht. Um jetzt zu einer Frage wie *Was trank Marie?* zu kommen, müssen wir zunächst die gesamte Phrase *einen Kaffee mit viel Milch* als intendierte Antwort identifizieren. (Um zu sehen, dass das nicht trivial ist, stelle man sich vor, der Satz hieße stattdessen *Marie trank einen Kaffee mit großer Begeisterung*.). Dann brauchen wir ein geeignetes Fragewort (*Was?*), müssen die Antwort aus dem Satz entfernen, das Fragewort passend einfügen und am Ende den Satz so umstellen, dass er den Regeln der deutschen Grammatik entspricht. Mit dieser Methode können wir zu einem beliebigen Text auf Deutsch oder Englisch große Mengen Fragen automatisch generieren, so dass Lehrende am Ende nur noch die Fragen auszuwählen brauchen, die ihnen geeignet erscheinen.

Multiple-Choice

Multiple-Choice-Aufgaben sind ein verbreitetes Aufgabenprinzip: Zu einer Frage oder einem Lückentext werden mehrere Antwortmöglichkeiten vorgegeben. Die Lernenden müssen die richtige Antwort aus einer Menge von falschen Antworten (den sogenannten Distraktoren) auswählen.

Während dieser Aufgabentyp sehr leicht automatisch ausgewertet werden kann, ist die automatische Erstellung vergleichsweise komplex: Man muss nicht nur eine Aufgabe generieren, sondern auch noch verschiedene, eindeutig richtige oder falsche Antwortmöglichkeiten vorgeben. Eine falsche Antwort zeichnet sich nicht alleine dadurch aus, dass sie falsch ist. Sie sollte zusätzlich zumindest potenziell plausibel erscheinen, im Idealfall müssen

die Lernenden länger nachdenken, um zu erkennen, welche Antwort richtig ist. Auf einer rein sprachlich-formalen Ebene bedeutet dies, dass die Antwort den richtigen Typ haben muss. Die Distraktoren für eine Wer-Frage sollten zum Beispiel Personen sein, die für eine Wo-Frage Ortsangaben. Wenn die Multiple-Choice-Aufgabe ein Lückentext ist, wie es bei Sprachlernübungen häufig vorkommt, müssen alle Antwortoptionen grammatikalisch in die Lücke passen.

Betrachten wir den Beispielsatz *Michael möchte ein altes Auto für 3000 Euro von seinem Freund Jan* _____. Eine mögliche korrekte Antwort wäre *übernehmen*. Wir müssen sicherstellen, dass alle Distraktoren, also alle inkorrekten Antworten, ebenfalls Verben sind, sonst wäre die Aufgabe für die meisten Lernenden zu einfach. Mögliche Distraktoren wäre also *trinken* oder *schreiben*. Kein guter Distraktor wäre *kaufen*: diese Antwort wäre annähernd synonym zu *übernehmen*. Solche Synonyme können durch statistische Analyse großer Textmengen automatisch ausgeschlossen werden. Schwierige falsche Antworten wären solche, die in bestimmten Kontexten synonym zur richtigen Antwort sind, aber im vorliegenden Kontext nicht passen. Ein solches Synonym wäre *weiterbeschäftigen (Die Firma hat alle Auszubildenden übernommen*). Im Kontext unseres Satzes passt der Begriff aber nicht, er ist eindeutig falsch. Solche Wörter sind schwierige und damit gute Distraktoren für eine Sprachlernaufgabe.

Gap-fill Bundles

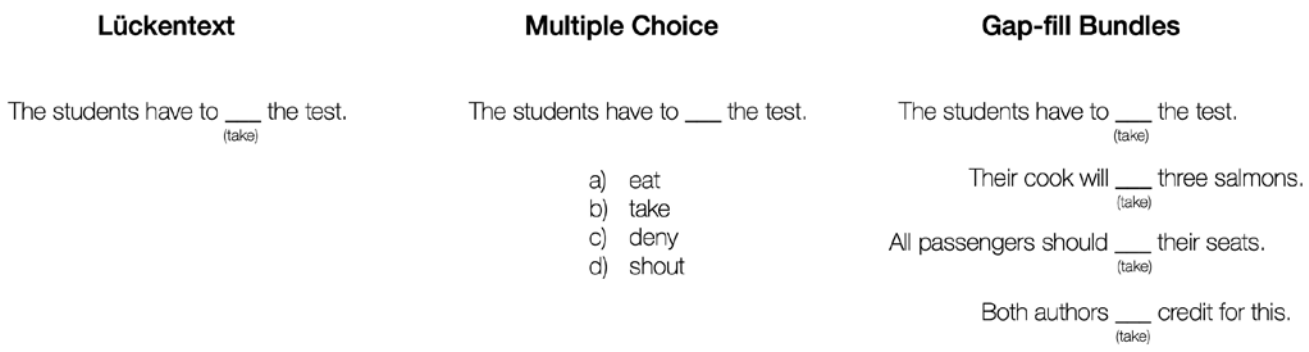
Gap-fill Bundles ähneln den oben erwähnten Lückentexten und Multiple-Choice-Aufgaben, vermeiden jedoch deren Defizite. Lückentexte sind leicht zu erstellen, aber vor allem, wenn ein ganzes Wort und nicht nur eine Wortendung eingefügt werden soll, sind einzelne Lücken fast immer mehrdeutig. Im Beispiel in Abbildung (2) links wären unter anderem *cook, buy* oder *eat* möglich. Multiple-Choice Aufgaben (im Bild in der Mitte) dagegen sind einfach zu korrigieren und, wie wir oben gesehen haben, genauso wie Lückentexte automatisch erstellbar. Sie sind allerdings aus didaktischer Sicht ungünstig, weil die Lernenden die richtige Antwort nur aus einer Menge von Möglichkeiten auswählen müssen, statt sie, wie bei einem Lückentext aktiv zu produzieren. Unser neues Format der gap-fill Bundles (in der Abbildung rechts) kombiniert die Vorteile beider Aufgabentypen und eliminiert die Nachteile. Die Lernenden sehen in einer Aufgabe ein Bundle von Lückentexten, bei denen in jede Lücke das gleiche Wort eingesetzt werden muss. Dadurch disambiguieren sich die Sätze gegenseitig, das Problem der Mehrdeutigkeit ist behoben. Gleichzeitig muss aktiv ein passendes Wort produziert und nicht nur ausgewählt werden.

Dieser Aufgabentyp ist für Lehrende nur mit sehr viel Zeitaufwand manuell zu erstellen, weswegen eine sprachtechnologische Lösung notwendig ist.[6]

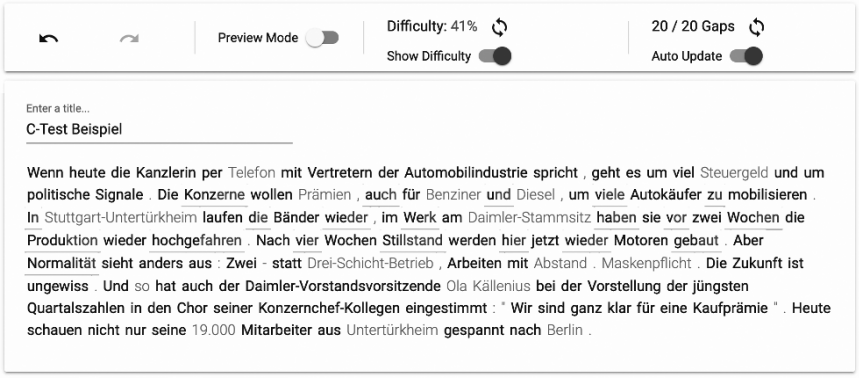
Zu einem gegebenen Zielwort werden aus einer Textsammlung Kandidatensätze ausgewählt. Man startet mit einem Satz und fügt dann weitere Sätze zu dem Bundle hinzu, so dass die Zahl der noch möglichen Antwortkandidaten möglichst gering wird, beziehungsweise nur eine Antwort über alle Sätze hinweg hinreichend wahrscheinlich ist. Dazu verwenden wir statistische Sprachmodelle, die vorhersagen können, wie wahrscheinlich ein Wort in einen bestimmten Satzkontext ist. Ein Bundle ist dann gut lösbar, wenn genau eine Lösung eine hohe, und alle anderen Lösungen eine sehr niedrige Wahrscheinlichkeit haben. Diese Lösungen können wir automatisch finden.

Sprachstandserhebung

Im Bereich des Sprachenlernens sind Lückentexte eine beliebte Möglichkeit, das aktuelle Sprachniveau zu ermitteln. Oft bedient man sich dabei sogenannter c-Tests, bei denen



(2) Beispiele für Aufgabenformate.
Quelle: eigene Darstellung



(3) Screenshot des c-Test-Builders.
Quelle: eigene Darstellung

in einem kurzen Text die zweite Hälfte jedes zweiten Wortes durch eine Lücke ersetzt wird. Solche Tests sind einfach automatisch auszuwerten: eine Lücke zählt als korrekt, wenn die erwartete Lösung eingesetzt wurde, sonst nicht, egal wie gering die Abweichung ist.

Die manuelle Erstellung eines gut balancierten c-Tests ist erneut ein zeitaufwändiger Prozess. Wir haben daher mit dem „c-Test-Builder" ein Werkzeug entwickelt, das Lehrende bei der Erstellung unterstützt. Abbildung (3) zeigt einen Screenshot des Werkzeuges mit einem automatisch erstellten Lückentext der nun noch manuell weiterbearbeitet werden kann.

Eine sprachtechnologische Herausforderung besteht unter anderem darin, Eigennamen, Jahreszahlen oder andere Zahlenangaben zu erkennen, da diese nicht in Lücken umgewandelt werden sollen, weil sie sehr schwer vorherzusagen sind. Insbesondere im Deutschen stellen auch Komposita ein Problem dar, da nicht das gesamte Wort, sondern nur der morphologische Kopf für die Lückengenerierung berücksichtigt wird. Würde man zum Beispiel beim Wort *Haustier* die komplette zweite Hälfte entfernen, erhielte man *Haus*_____ und damit eine sehr mehrdeutige Lücke (Hausarzt, Haushalt, Hauskauf, Hauswand, ...). Stattdessen erkennen wir *tier* automatisch als Kopf des Kompositums und reduzieren das Wort zu *Hausti*_____ .

Ein nach diesen Regeln erstellter c-Test entspricht den formalen Richtlinien, kann aber für die Lernstandserhebung immer noch ungeeignet, da zu leicht oder schwer, sein. Da viele Faktoren die Schwierigkeit eines c-Tests beeinflussen, ist die Vorhersage auch für Lehrende sehr schwer. Stattdessen kann ein auf maschinellem Lernen basierender Ansatz genutzt werden, der die Schwierigkeit jeder Lücke individuell vorhersagen kann.[1]

Das von uns entwickelte c-Test-Builder-Werkzeug kennt die Besonderheiten von verschiedenen Sprachen und ermöglicht es Lehrenden so, schnell und einfach c-Tests auf dem richtigen Schwierigkeitsniveau zu erstellen und so zu exportieren, dass diese beispielsweise auch direkt in die verschiedenen Lernplattformen importiert werden können.

Bewertung

Nachdem eine Aufgabe von den Lernenden bearbeitet wurde, muss üblicherweise das Ergebnis bewertet werden. Dies gilt selbst für reine Übungsaufgaben, deren Wirksamkeit deutlich erhöht wird, wenn die Lernenden auch erfahren, ob sie richtig geantwortet haben. Gerade in der digitalen Lehre mit potenziell sehr vielen Teilnehmenden ist eine Unterstützung der Lehrenden bei der Bewertung vorteilhaft.

Eine automatische Auswertung kann auf verschiedene Arten genutzt werden. Im einfachsten Fall wird nur zurückgemeldet, ob eine Antwort richtig oder falsch ist, beziehungsweise wie viele Punkte erreicht wurden. Wie diese Beurteilung zustande kommt, diskutieren wir hier.

Natürlich kann, und sollte im Idealfall, den Lerner*innen nach Möglichkeit ein weitergehendes Feedback gegeben werden, dazu mehr unter „Feedback".

Bei bestimmten Aufgabentypen ist eine maschinelle Unterscheidung in richtig oder falsch sehr einfach, zum Beispiel bei Multiple-Choice-Aufgaben, bei denen nur geprüft werden muss, welche Antwort ausgewählt wurde.

Bei Aufgabenformaten mit Freitextantworten ist diese Aufgabe deutlich anspruchsvoller.

Betrachten wir folgendes Beispiel: Anna soll in einer Physikaufgabe entscheiden, was mit einer bestimmten Glühbirne (*Birne A*) in einem Schaltplan passiert, wenn eine andere Birne (*Birne B*) durchbrennt. Eine Musterantwort könnte lauten *Birne A leuchtet weiter, weil sie zu Birne B parallel geschaltet ist.*

Ein Algorithmus sollte aber auch erkennen, dass eine Antwort wie *Birne A bleibt an, weil ihr Stromkreislauf mit dem Durchbrennen von B nicht unterbrochen wird* auch richtig ist, obwohl sie oberflächlich wenig mit der Musterantwort gemeinsam hat. Gleichzeitig sollte *Birne A leuchtet nicht weiter, weil sie zu Birne B seriell geschaltet ist.* trotz oberflächlich hoher Übereinstimmung als falsch bewertet werden. Die Herausforderung besteht also generell in der Varianz der möglichen Antworten. Diese manifestiert sich in unterschiedlichen sprachlichen Realisierungen, also verschiedenen Formulierungen, von denen einige zwar sprachlich, also zum

Beispiel orthographisch, falsch, aber inhaltlich immer noch korrekt sein können.

Je nach Aufgabentyp können verschiedene Maßstäbe zur Bewertung angelegt werden. Im Bereich des automatischen Scorings unterscheiden wir vor allem zwischen zwei Arten von Freitextantworten: Aufsätze und Kurzantworten.

Bei einem **Aufsatz** werden längere Text zu einer bestimmten Fragestellung geschrieben, zum Beispiel *Halten Sie Studiengebühren für gerechtfertigt?*. Solche Aufsätze werden einerseits nach inhaltlichen Kriterien bewertet. Werden für das Thema relevante Argumente vorgebracht und durch plausible Beispiele untermauert? Andererseits wird die Sprache bewertet, also Rechtschreibung und Grammatik, das benutzte Vokabular, der Aufbau und so weiter. Im Normalfall kann man weder eine gute Note bekommen, wenn man sprachlich eloquent, aber am Thema vorbeischreibt, noch wenn man inhaltlich sinnvoll aber sprachlich mangelhaft schreibt. Solche Aufsätze dienen oft dazu, den Sprachstand eines Lerners, zum Beispiel in einer Fremdsprache, zu bestimmen.

Bei **Kurzantworten** sieht die Sache anders aus. Hier werden Faktenfragen gestellt, wie zum Beispiel *Wann wurde Mozart geboren?*, *Wie definiert sich eine monoton steigende Funktion?* oder *Wie funktioniert Photosynthese?.* Bei der Bewertung der Antworten geht es einzig und allein um die inhaltliche Korrektheit. Wie sprachlich schön oder orthographisch korrekt die Antwort geschrieben ist, sollte keine Rolle spielen.

Aus diesen Unterschieden leiten sich auch unterschiedliche Erfordernisse für die automatische Bewertung ab.

Grundsätzlich nutzen wir bei der automatischen Bewertung Methoden des maschinellen Lernens. Das bedeutet, wir trainieren ein Computermodell mit Beispielen von Lernerantworten und vom Menschen vorgegebenen Bewertungen. Aus diesen Beispielen lernt das Modell, wie ein Mensch eine Antwort bewerten würde und kann diesen Algorithmus dann auf neue, unbewertete Antworten übertragen. Um Lernertexte überhaupt verarbeiten zu können, werden sie zunächst durch Merkmale repräsentiert, wobei Aufsätze und Kurzantworten sich teilweise deutlich voneinander unterscheiden. Bei beiden ist der Inhalt des Geschriebenen wichtig. Eine Standardrepräsentation besteht darin die verwendeten Wörter und Wortkombinationen zu betrachten. Bei der Frage *Wo wurde Mozart geboren* dürfte beispielsweise das Vorhandensein des Wortes *Salzburg* ein wesentliches Merkmal für die Korrektheit der Antwort sein. Je nachdem, welcher Algorithmus für das maschinelle Lernen benutzt wird, gibt es unterschiedliche Arten, dieses Merkmal einzusetzen. In vielen Verfahren nutzt man das Vorhandensein oder nicht eines bestimmten Wortes als binäres Merkmal. Man kann aber ein Wort auch selbst durch einen Merkmalsvektor (ein sogenanntes „Embedding“) repräsentieren und so dessen Bedeutung nuancierter ausdrücken. Der Merkmalsvektor beschreibt, in welchem Kontexten dieses Wort vorkommt. Dadurch ähneln sich beispielsweise die Vektoren für *Stuhl* und *Hocker*, so dass ein Algorithmus die Antwort *Peter sitzt auf einem Hocker* mit einer gewissen Wahrscheinlichkeit auch als richtig bewertet, selbst wenn im Training nur die Antwort *Peter sitzt auf einem Stuhl* vorkam.

Bei der Bewertung von Aufsätzen nutzen wir darüber hinaus auch eine Vielfalt von linguistischen Features, die die sprachliche Qualität eines Aufsatzes modellieren. Dies sind zum Beispiel die Anzahl und Art der Rechtschreib- und Grammatikfehler in einem Text, die Häufigkeit von bestimmten grammatischen Konstruktionen, wie Nebensätze, Passiv oder Substantivierungen, oder die lexikalische Varianz und Häufigkeit des verwendeten Vokabulars.

Mit all diesen Bewertungsmethoden können wir beispielsweise Tests zur Sprachstandsdiagnostik im Fremdsprachenunterricht oder Aufsätze, die von Lehramtskandidaten geschrieben wurden, automatisch bewerten. Kombiniert man solche Methoden mit automatischer Handschrifterkennung, können wir sogar handgeschriebene Klausuren mit dem Computer verarbeiten.

Unsere aktuelle Forschung beschäftigt sich darüber hinaus mit der Frage, wie robust die Verfahren gegenüber Betrugsversuchen unter anderem durch vom Computer generierte Texte sind. Bestehende Verfahren weisen solchen Aufsätzen überraschend oft eine gute Note zu.

Auch die Frage, in welchen Sprachen wir Texte automatisch bewerten können, beschäftigt uns. Die meisten Modelle arbeiten auf deutschen oder englischen Lernerdaten, aber wir haben beispielsweise auch Daten auf Spanisch und Chinesisch erhoben und untersuchen, ob sich Modelle von einer in eine andere Sprache übertragen lassen.

Feedback

Ein wichtiger Faktor im Lernprozess ist für Lernende das Feedback, das sie zu abgeschlossenen Aufgaben erhalten. Natürlich ist die reine Beurteilung in Form von „richtig/falsch” oder abgestuften Varianten davon im Sinne einer Noten- oder Punkteskala bereits eine Form von Rückmeldung über den Aufgabenerfolg. Allerdings liefert dies in den seltensten Fällen Informationen dazu, was man eigentlich falsch gemacht hat beziehungsweise was in Zukunft anders gemacht werden müsste.

Für einige Aufgabentypen mag es ausreichend sein, eine Musterantwort zur Verfügung zu stellen, zum Beispiel bei Kurzantworten. Gerade bei kurzen Faktenfragen ist das oft ausreichend. Wenn jemand auf die Frage *Wo wurde Mozart geboren? Wien* antwortet und dann erfährt, dass die richtige Antwort *Salzburg* gewesen wäre, ist das normalerweise informativ genug. Das System sollte nicht

zusätzlich noch Selbstverständlichkeiten erklären, wie beispielsweise, dass eine Person nicht an zwei Orten gleichzeitig geboren sein kann, und deshalb *Wien* falsch ist.

Bei etwas komplexeren Fragen bleibt es bei diesem Vorgehen jedoch noch immer den Lernenden überlassen, zu erkennen, inwiefern eine Antwort von der Musterantwort abweicht.

Nehmen wir als Beispiel die Physikaufgabe von oben, bei der entschieden werden soll, was mit Birne A in einem Schaltplan passiert, wenn Birne B durchbrennt. Angenommen jemand antwortet *Birne A geht aus, weil sie mit Birne B in Reihe geschaltet ist.*, dann könnte eine gute Rückmeldung folgendermaßen aussehen: *Birne A und Birne B sind parallel geschaltet und nicht in Reihe, wie Sie geschrieben haben. Deshalb geht Birne A auch nicht aus, sondern leuchtet weiter. Die Antwort ist daher falsch.* Hier steht die Forschung noch ziemlich am Anfang. Vielversprechende Ansätze lernen automatisch zu einem Paar aus Lerner- und Musterantwort eine Erklärung zu generieren, die die logische Beziehung zwischen den beiden Sätzen verbalisiert. Ist die Antwort zum Beispiel falsch, weil sie der Musterantwort widerspricht, oder weil wichtige Details fehlen?

Besonders groß ist die Herausforderung, wenn Aufsätze automatisch bewertet werden. Hier „versteht" der Computer den Inhalt noch nicht auf einer Ebene, die ihm erlauben würde, zu beurteilen, ob die Argumente schlüssig sind oder die Sätze kohärent miteinander verbunden wurden.

In einigen Bereichen sind ausführliche Rückmeldungen vom Zeitaufwand her aber auch kaum von Menschen zu leisten. Einer dieser Bereiche ist der Orthographieerwerb von Grundschulkindern. Um die Fehlerschwerpunkte eines Kindes zu ermitteln und darauf basierend zum Beispiel passendes Übungsmaterial bereitzustellen, ist es notwendig, einzelne Fehlertypen voneinander zu unterscheiden. Das ist im Deutschen mitunter gar nicht so einfach, wie die folgenden zwei Fehlschreibungen deutlich machen:

(a) **tenken* für *denken*
(b) **Hunt* für *Hund*

In beiden Fällen wurde ein <t> für ein <d> geschrieben, doch die Fehler sind ganz unterschiedlicher Art: Während man bei **tenken* einen lautlichen Unterschied zu *denken* wahrnimmt, ist dies bei **Hunt* für *Hund* nicht der Fall. Auf Grund des Phänomens der Auslautverhärtung spricht man im Deutschen das <d> am Ende eines Wortes wie ein <t>. Somit wäre es hier für das Kind vollkommen irreführend, darauf hinzuweisen, dass es sich das Wort deutlicher vorsprechen müsse. Jeden einzelnen Rechtschreibfehler in einem Text auf diese Art zu klassifizieren, ist jedoch sehr mühsam und zeitaufwendig. Ein Instrument, das für genau eine solche manuelle Fehlerklassifikation entwickelt wurde, ist beispielsweise die Oldenburger Fehleranalyse (OLFA). Sie stellt 37 unterschiedliche Fehlerkategorien zur Verfügung und gibt an, dass man bei einer Textmenge von etwa 350 Wörtern einen Zeitbedarf von über 30 Minuten für das Markieren der Fehler und noch einmal mindestens 30 Minuten für die Zuordnung in die Fehlerkategorien einkalkulieren sollte. Es ist klar, dass dies innerhalb des normalen Schulbetriebs praktisch nicht zu leisten ist. Wir arbeiten daher an einer automatischen Rechtschreibfehlerklassifikation. Dazu wird ein Wort auf mehreren linguistischen Ebenen analysiert (z.B. Aussprache, Wortbestandteile) um die korrekten Fehlerkategorien zuordnen zu können. Die automatische Klassifikation ermöglicht Lehrenden dann, auf einen Blick die Fehlerschwerpunkte eines Kindes zu erkennen.[4]

Ethische Gesichtspunkte

Der Einsatz von Technologie, insbesondere künstlicher Intelligenz, kann bestimmte ethische Fragestellungen aufwerfen. So besteht zum Beispiel die Gefahr, dass ein Algorithmus bestimmte Gruppen systematisch benachteiligt. Wenn beispielsweise eine automatische Bewertungsmethode Frauen grundsätzlich besser oder schlechter bewerten würde als Männer, wäre das inakzeptabel. Automatische Systeme können solche fragwürdigen Unterscheidungen auch anhand von impliziten Merkmalen hervorrufen, zum Beispiel falls Männer mehr Rechtschreibfehler machen würden als Frauen. Betrachten wir als Beispiel nochmal die oben besprochenen Kurzantwortfragen, bei denen es ja für die Bewertung nur auf den Inhalt, aber nicht auf die sprachliche Korrektheit ankommen darf. Gleichzeitig wissen wir, dass viele automatische Sprachverarbeitungswerkzeuge durchaus ein Problem mit fehlerhaftem Input haben und einen falsch geschriebenen Satz schlechter analysieren können als mustergültiges Zeitungsdeutsch. Daraus könnte eine schlechtere Bewertung von Antworten mit vielen Fehlern folgen. Wir konnten in einer Studie zeigen, dass aktuelle Verfahren zur automatischen Bewertung bei einer moderaten Menge an Fehlern immer noch zu einer korrekten Bewertung kommen. Erst bei ungewöhnlich vielen Fehlern bricht die Leistungsfähigkeit ein, allerdings wäre noch zu prüfen, ob bei so vielen Fehlern nicht auch die Verständlichkeit in einem Maß leidet, das auch bei manueller Bewertung eine Abwertung gerechtfertigt hätte. Natürlich braucht es darüber hinaus die systematische Untersuchung der Bewertungsvorgänge, ob nicht weitere Merkmale zu einer ungerechten Beurteilung führen könnten.

Grundsätzlich besteht natürlich auch immer die Möglichkeit, automatische Bewertung nicht vollautomatisch, sondern im Rahmen einer Bewertungsunterstützung anzuwenden. In solchen Fällen erhalten Lehrende beispielsweise Bewertungsvorschläge vom Computer, die

Andrea Horbach. Foto: Vladimir Unkovic

Ronja Laarmann-Quante. Foto: Vladimir Unkovic

sie immer noch akzeptieren oder zurückweisen müssen oder bekommen Gruppen von inhaltlich ähnlichen Antworten zur gemeinsamen Bewertung angezeigt. Durch solche Methoden entlastet man Lehrende, ohne sie komplett aus dem Bewertungsprozess auszuschließen.

Der Einsatz von automatisierten Systemen kann sich natürlich auch positiv auswirken. Ein einmal trainiertes Verfahren bewertet alle einkommenden Antworten ohne Ansehen der Person gleich und ohne Verzögerung und kann damit insbesondere bisher benachteiligten Personen den Zugang zu Bildungsangeboten deutlich erleichtern.

Zusammenfassung

Die anspruchsvolle und zeitaufwändige Aufgabe der Wissensvermittlung kann durch Methoden der künstlichen Intelligenz, insbesondere der automatischen Sprachverarbeitung, unterstützt werden.

In diesem Artikel konzentrieren wir uns auf drei Hauptanwendungsgebiete: Wir generieren automatisch Übungen für verschiedene Fragetypen, die für Menschen schwer oder mühsam zu formulieren sind, und beurteilen den Schwierigkeitsgrad der von uns erstellten Übungen. Wir beurteilen automatisch die Antworten der Lernenden, seien es inhaltsbezogene Kurzantworten, die auf der Grundlage ihres Inhalts bewertet werden, oder längere Aufsätze, die eher auf der Grundlage des Sprachgebrauchs und der Argumentation bewertet werden. Schließlich geben wir den Lernenden eine Rückmeldung über ihren Text, zum Beispiel indem wir die von Schulkindern produzierten Rechtschreibfehler analysieren, so dass sie eine Rückmeldung darüber erhalten, mit welchen orthographischen Phänomenen sie noch Schwierigkeiten haben.

Die meisten sprachtechnologischen Methoden beruhen darauf, dass wir aus vorliegenden Daten lernen. Wir sind daher auf die Zusammenarbeit mit Lehrenden angewiesen. Wenn Sie ein interessantes Problem haben, dass zu den hier diskutierten Themen passt, sprechen Sie uns gern an.

Summary

Artificial intelligence, especially methods of natural language processing, can be used in educational contexts in order to support the teaching process. In this article, we focus on three main application types that we are currently working on: we automatically **generate** exercises for various question types that are hard or tedious to formulate for humans, and assess the difficulty of exercises we created. We automatically **score learner answers**, be it answers to short answer questions, which are scored based on their content, or longer essays, scored based on language use and argumentation quality. Finally we **give learners feedback** about their texts, for example by analyzing the spelling errors produced by school children in order to inform them about the orthographic phenomena they still struggle with.

Anmerkungen/Literatur

1) Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. Transactions of the Association for Computational Linguistics (TACL), 2, 517–530.
2) Horbach, A., Scholten-Akoun, D., Ding, Y., & Zesch, T. (2017, September). Fine-grained essay scoring of a complex writing task for native speakers. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 357–366).
3) Horbach, A., & Zesch, T. (2019). The Influence of Variance in Learner Answers on Automatic Content Scoring. In Frontiers in Education (Vol. 4, p. 28).
4) Laarmann-Quante, R. (2017). Towards a Tool for Automatic Spelling Error Analysis and Feedback Generation for Freely Written German Texts Produced by Primary School Children. In SLaTE (pp. 36–41).
5) Meurers, D. (2012). Natural language processing and language learning. The Encyclopedia of Applied Linguistics.
6) Wojatzki, M., Melamud, O., & Zesch, T. (2016, June). Bundled gap filling: A new paradigm for unambiguous cloze exercises. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 172–181).

*Die Autor*innen*

Torsten Zesch studierte von 1999 bis 2005 Informatik an der Technischen Universität Chemnitz. Er promovierte von 2006 bis 2009 an der Technischen Universität Darmstadt und leitete dort von 2010 bis 2013 die Nachwuchsgruppe „Intelligente Sprachsysteme“. Nach Stationen als Visiting Researcher in Israel und den USA sowie als Vertretungsprofessor am Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF) kam er 2014 als Juniorprofessor für Sprachtechnologie an die Universität Duisburg-Essen. 2020 erfolgte die Ernennung zum Universitätsprofessor. Torsten Zesch ist Erster Vorsitzender der „Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)“.

Andrea Horbach studierte von 2002 bis 2007 Computerlinguistik und von 2008 bis 2009 Deutsch als Fremdsprache an der Universität des Saarlandes. Von 2010 bis 2016 war sie wissenschaftliche Mitarbeiterin am Institut für Computerlinguistik in Saarbrücken und arbeitet seit 2016 als wissenschaftliche Mitarbeiterin am Lehrstuhl für Sprachtechnologie an der Universität Duisburg-Essen in Forschungsprojekten zu Sprachverarbeitungsmethoden im Bildungsbereich. Ihre Promotion schloss sie 2018 zum Thema der automatischen Bewertung von Freitextantworten ab.

Ronja Laarmann-Quante studierte von 2010 bis 2015 Linguistik mit Schwerpunkt Computerlinguistik an der Ruhr-Universität Bochum. Von 2015 bis 2019 arbeitete sie dort als wissenschaftliche Mitarbeiterin in einem interdisziplinären Projekt zum Schriftspracherwerb von Kindern und promovierte angelehnt daran 2020 zum Thema der automatischen Vorhersage von Rechtschreibfehlern. Seit Mitte 2019 ist sie als wissenschaftliche Mitarbeiterin am Lehrstuhl für Sprachtechnologie an der Universität Duisburg-Essen tätig.