Sonderdruck aus: Diagnostica, 53, Heft 3, 155–165

Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse

Wolfgang Lenhard, Herbert Baier, Joachim Hoffmann und Wolfgang Schneider

Zusammenfassung. Das Schreiben von Kurzaufsätzen im Rahmen informeller Diagnostik ist weit verbreitet, jedoch mit Problemen der Auswertungsobjektivität behaftet. Durch die computerbasierte Bewertung von Essays lassen sich Entscheidungsheuristiken vermeiden. Die latente semantische Analyse (LSA) ist ein statistisches Verfahren, dass die Repräsentation von Wissensstrukturen im Computer und somit die automatische inhaltliche Bewertung von Aufsätzen ermöglicht. Es wird die Entwicklung eines deutschsprachigen LSA-basierten Systems beschrieben und seine Leistungsfähigkeit in einem Multiple-Choice-Wissenstest, sowie bei der Bewertung von Aufsätzen aufgezeigt. Bei der Klassifikation von Tierarten nach Klassen erzielte es vergleichbare Leistungen wie Studierende. Bei der Bewertung von Aufsätzen wurden Korrelationen mit menschlichen Bewertern im mittleren bis oberen Bereich erzielt. Die Übereinstimmung von Summenscores mehrerer Aufgaben reicht an für standardisierte Verfahren geforderte Reliabilitätskennwerte heran. Neben der automatischen Aufsatzbeantwortung liegen Anwendungen v. a. in intelligenter Lernsoftware und der Ergänzung bestehender psychologischer Modelle durch semantische Module.
Schlüsselwörter: latente semantische Analyse, LSA, automatische Aufsatzbewertung, automatische Kategorisierung

Automatic scoring of constructed-response items with latent semantic analysis

Abstract. The validity of constructed-response items like essays that are commonly used within informal diagnostics is threatened by many biases. Computer based essay scoring reduces these biases. Latent semantic analysis (LSA) is a statistical technique that allows the representation of human semantic knowledge structures within the computer, thus enabling automatic essay scoring. This article describes the basic steps for implementing a LSA-based system in German language and examines its performance on multiple-choice knowledge tests and essay scoring. It showed an equal performance compared to university students in classifying animal species and achieved medium to high correlations with human raters in essay scoring. Using cumulated scores, the correlations reached values necessary for the reliability of standardized tests. Besides essay scoring, intelligent tutoring software and the extension of psychological models by semantic modules are further interesting fields of application for LSA.
Key words: latent semantic analysis, LSA, automated essay scoring, automatic categorization

Das Schreiben von Kurzaufsätzen gehört zu den am weitesten verbreiteten und höchst geschätzten Formen der Überprüfung des Wissens und der Argumentationsfähigkeiten von Schülern und Studenten (Miller, 2003). Dieses Antwortformat nimmt im Rahmen informeller Diagnostik, beispielsweise bei der Notenvergabe oder in Studieneingangstests, einen hohen Stellenwert ein. Ein Grund dafür liegt darin, dass bei offenen Antworten komplexere Aufgabenstellungen konstruierbar sind und im Gegensatz zu Multiple-Choice-Tests Wissen frei reproduziert werden muss. Auch ist die fundierte Konstruktion eines Multiple-Choice-Tests wesentlich arbeitsintensiver als das Stellen von Essay-Fragen. Im Gegensatz dazu gilt ihre Verwendung in standardisierten Verfahren als problematisch (vgl. Lienert & Ratz, 1998, 21 f.), da die Auswertung aufwändig ist und Objektivitätsprobleme birgt. So muss i. d. R. eine Musterlösung oder ein eindeutiges Bewertungsschema erstellt werden, wobei sich die Klassifikation der Einzelaussagen eines Kurzaufsatzes als richtig oder falsch dann dennoch meist als schwierig herausstellt (Bühner, 2004, 60 f.). Die größte Fehlerquelle in Bezug auf die Auswertungsobjektivität liegt aber vermutlich im Bewerter selbst. Es konnte eine große Anzahl an verzerrenden Entscheidungsheuristiken identifiziert werden, die die Auswertungsobjektivität negativ beeinflussen (vgl. Haladyna,

Das Forschungsprojekt wird aus Mitteln der deutschen Forschungsgemeinschaft finanziert (Förderkennzeichen: HO 1301/11-2 & SCHN 315/29-1). Wir danken der LSA-Research-Group und insbesondere Profs. Drs. Walter und Eileen Kintsch, Prof. Dr. Thomas Landauer (University of Boulder/Colorado), sowie Prof. Dr. Guy Denhière und Prof. Dr. Sandra Jhean-Larose (Paris) für ihre Unterstützung. Für die Bereitstellung von Texten in elektronischer Form danken wir Prof. Dr. Wolfgang Schönpflug und Dr. Ute Schönpflug und den Verlagen Spektrum Akademischer Verlag und Springer. Das Forschungsprojekt ist unter der Adresse http://www.summa.psychologie.uni-wuerzburg.de näher dargestellt.

DOI: 10.1026/0012-1924.53.3.155

1999, 43 f.), darunter die Schönheit der Handschrift (Chase, 1979; Marshall & Powers, 1969), Länge der Sätze (Coffmann, 1971), Reihenfolgeeffekte bei der Bewertung (Hughes et al., 1983), Themenwahl und Vergleich von Bewertungen für Essays zu verschiedenen Themen (Meyer, 1939), sowie Geschlecht und ethnische Zugehörigkeit des Schreibers oder der Schreiberin (Chase, 1986). Die Überwindung dieser Mängel mittels automatischer Bewertung von Aufsätzen war in den letzten Jahren das Ziel großer Forschungsanstrengungen vorwiegend im englischsprachigen Raum (Page, 1966; Burstein, Kukich, Wolff, Lu & Chodorow, 1998; Landauer, Laham, Rehder & Schreiner, 1997; vgl. Ishioka & Kameda, 2006; Miller, 2003). Wesentlich stimuliert wurde die Forschung durch die Entwicklung von statistischen Verfahren, die eine Simulation von Teilaspekten semantischen Wissens und verbaler Intelligenz am Computer ermöglichen.

Im Folgenden wird dargestellt, welche Schritte für die automatische Bewertung von Aufsätzen im Deutschen notwendig sind und welche Höhe der Übereinstimmung mit menschlichen Bewertern erzielt werden können.

Computerbasierter Vergleich von Textinhalten mittels latenter semantischer Analyse

Damit ein Computer Texte inhaltlich bewerten kann, müsste er über ein ausreichendes Maß an verbaler Intelligenz und über semantisches Wissen verfügen. Dies ist ein bislang ungelöstes Problem der Forschung im Bereich künstlicher Intelligenz, jedoch wurden verschiedene Verfahren entwickelt, deren Ziel die Repräsentation von Textinhalten am Computer ist (vgl. Lemaire & Denhière, 2004; Landauer, McNamara, Dennis & W. Kintsch, in press). Das prominenteste dieser Verfahren trägt die Bezeichnung latente semantische Analyse (LSA) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). Dabei handelt es sich um eine Technik der automatischen Sprachverarbeitung, die in Bezug auf den inhaltlichen Vergleich von Textinhalten Teilaspekte von Wortbedeutungen und semantischen Wissens hinreichend gut simulieren kann. Sie ermöglicht die Analyse der Beziehung zwischen Wörtern auf der Basis ihres gemeinsamen Auftretens. Es handelt sich um einen rein statistischen Ansatz, d. h., die auf Kookurrenzen basierenden Wortbeziehungen werden automatisch extrahiert, ohne dass vorab Regelsysteme spezifiziert oder Wörterbücher eingegeben werden müssen. Das Verfahren weist Parallelen zur Faktorenanalyse auf, weswegen zum besseren Verständnis im Folgenden Bezüge zwischen beiden Ansätzen aufgezeigt werden.

Generierung semantischer Räume

Ausgangsbasis der LSA sind Textsammlungen, wobei das Textmaterial üblicherweise in Absätze aufgespaltet wird (im Folgenden als Dokument bezeichnet). Die in den Dokumenten gespeicherten Informationen über Wortbeziehungen lassen sich in einer Frequenzmatrix abstrakt repräsentieren, wobei die Spalten die einzelnen Dokumente und die Zeilen die unterschiedlichen Wörter umfassen. In den Zellen findet sich die Auftretenshäufigkeit eines Wortes in einem bestimmten Dokument. Verwendet man große Korpora natürlicher Sprache, dann ist diese Frequenzmatrix sehr dünn besetzt. Im Deutschen weisen beispielsweise mehr als 99.9 % der Zellen als Wert eine 0 auf. Die Frequenzmatrix enthält bereits sämtliche Informationen über Wortbeziehungen. Sie ist aber in der Regel zu groß, um damit Berechnungen durchführen zu können und sie enthält zum großen Teil unnötige Informationen („Störgeräusche“). Um die „Störgeräusche“ zu eliminieren und die in der Frequenzmatrix enthaltenen Information auf den Kerngehalt zu reduzieren sind vier Schritte notwendig: Filterung potenziell überflüssiger Wörter, Anwendung von Gewichtungsfunktionen auf die Zellhäufigkeiten, Singulärwertzerlegung und Bestimmung der optimalen Anzahl an Dimensionen.

Im ersten Schritt werden potenziell überflüssige Wörter ausgeschlossen. Hierzu gehören hochfrequente Wörter, die keine spezifische Information transportieren („Stopp-Wörter“, z. B. Präpositionen, Konjunktionen, Artikel usw.), sowie Wörter, die sehr selten auftreten, beispielsweise weniger als drei Mal im gesamten Textkorpus. Hierdurch reduziert sich die Anzahl unterschiedlicher Wörter deutlich. Als nächstes wird auf die Zellhäufigkeiten eine Gewichtungsfunktion angewendet. Die besten Resultate erbringt vermutlich die so genannte Log-Entropie-Gewichtung (Nakov, Popova & Mateev, 2001), bei der Zellhäufigkeiten hervorgehoben werden, wenn ein Wort lokal gehäuft vorkommt und somit spezifisch für einen bestimmten Kontext ist. Gleich häufig vorkommende Wörter, die aber sehr verteilt auftreten, werden hingegen abgeschwächt, da sie keine spezifische Information transportieren. Als letztes wird die gefilterte und gewichtete Matrix einer Singulärwertzerlegung unterzogen, ähnlich wie dies bei einer Faktorenanalyse der Fall ist. Während bei einer Faktorenanalyse die Kovarianzmatrix zerlegt wird, nimmt man bei der LSA die Matrix mit den gewichteten Auftretenshäufigkeiten. In beiden Fällen entstehen zwei orthogonale Matrizen und eine Diagonalmatrix, die durch Multiplikation wieder die Ausgangsmatrix ergeben. Wird eine Faktorenanalyse auf die verschiedenen Variablen einer Stichprobe von Objekten angewandt, so entsteht eine Matrix mit den Faktorwerten der einzelnen Objekte auf den verschiedenen extrahierten Dimensionen (Faktoren). Weiterhin eine Diagonalmatrix mit den sortierten Eigenwerten der Dimensionen und eine Matrix der Faktorladung der Variablen. In der LSA entstehen analog eine Wortmatrix mit den Faktorwerten der Wörter, eine Matrix der sortierten Singulärwerte und eine Dokumentmatrix (vgl. Abb. 1, mathematische Beschreibung siehe Berry, Dumais & O’Brien, 1995; Martin & Berry, in press).

Durch die Reduktion der Anzahl der Dimensionen der Teilmatrizen gehen „Störgeräusche“ der Rohmatrix verloren und man erhält ähnlich wie in der Faktorenanalyse eine Einfachstruktur mit den wesentlichen Informationen der

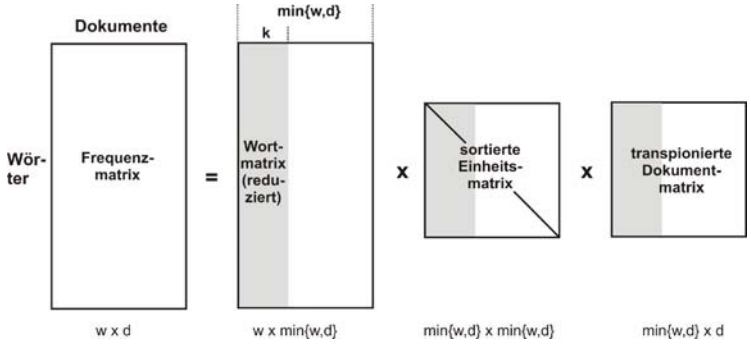

Abbildung 1. Bei der Singulärwertzerlegung entstehen drei Teilmatrizen, die durch Multiplikation die ursprüngliche Matrix ergeben. Für Ähnlichkeitsberechnungen im Rahmen der LSA wird die reduzierte Wortmatrix und die reduzierte sortierte Einheitsmatrix verwendet.

Ausgangsmatrix. Da die Singulärwertzerlegung in der LSA auf gewichteten Rohwerten basiert, ist die Höhe der Singulärwerte allerdings nicht interpretierbar, sodass a priori kein sinnvolles Kriterium für die Anzahl an Dimensionen festgelegt werden kann. Werte um die 300 Dimensionen haben sich als optimal erwiesen, wobei die Empfehlungen verschiedener Autoren zwischen 100 und 1500 Dimensionen variieren (Dumais, 1990; Graesser et al., 1999; Nakov, 2000; Wild, Stahl, Stermsek & Neumann, 2005).

Durch die Dimensionsreduktion wird letztlich ein Raum etabliert, in dem Wörter nach ihrem gemeinsamen Vorkommen mit anderen Wörtern verteilt sind. Der „Ort" eines Wortes im Raum (sein Vektor) repräsentiert damit den Teil seines Inhaltes, der sich in seiner gemeinsamen Verwendung mit anderen Wörtern manifestiert (deswegen auch *Latente* Semantische Analyse). Dementsprechend werden nicht nur Wörter benachbart repräsentiert, die oft gemeinsam verwendet werden, sondern auch Wörter, die niemals direkt zusammen aber oft gemeinsam mit gleichen anderen Wörtern verwendet werden (Zusammenhänge höherer Ordnung, vgl. Lemaire & Denhière, 2004; Kontostathis & Pottenger, 2002). Dies trifft beispielsweise auf Synonyme zu, die meist nicht gemeinsam auftreten, normalerweise jedoch im gleichen Kontext verwendet werden. Gleiches gilt für Ein- und Mehrzahl von Substantiven und Flexionen von Verben.

Im englischsprachigen Bereich hat sich als Bezeichnung für die in den reduzierten Matrizen enthaltenen Daten der Begriff *semantischer Raum* (semantic space) eingebürgert, sodass im Folgenden diese Terminologie übernommen wird. Ein semantischer Raum kann als n-dimensionale Vektorrepräsentation von Wort- und Textbedeutungen angesehen werden: Jedes Wort erhält durch seine Faktorwerte eine Koordinate in diesem Raum und somit einen Vektor mit einer bestimmten Richtung und Länge. Die Richtung ist dabei ein Analogon zur Thematik des Wortes, wohingegen die Länge des Vektors seinen semantischen Gehalt widerspiegelt.

Berechnung der semantischen Ähnlichkeit von Wörtern und Texten

Je nach Anzahl der Dimensionen kann die Berechnung des semantischen Raumes eine große Menge Arbeitsspeichers und eine lange Rechenzeit erfordern. Dagegen ist eine Ähnlichkeitsberechnung auf der Basis eines einmalig generierten semantischen Raumes in wenigen ms möglich. Eine solche Ähnlichkeitsberechnung ist durch den Vergleich des Zwischenwinkels, der Vektorlängen oder auch des euklidischen Abstandes der Koordinaten möglich. Daneben sind viele weitere Distanz- und Ähnlichkeitsmaße, sowie Kombinationen denkbar. Die Berücksichtigung der Vektorlänge beim inhaltlichen Vergleich von Textmaterial erbringt gegenüber der ausschließlichen Betrachtung des Zwischenwinkels allerdings oftmals keine besseren Ergebnisse (Landauer, Laham, Rehder & Schreiner, 1997; Rehder, Schreiner, Wolfe, Laham, Landauer & Kintsch, 1998). Der Kosinus des Zwischenwinkels kann zudem sehr einfach wie eine lineare Korrelation interpretiert werden.

Möchte man nun die Ähnlichkeit von Textinhalten berechnen, so müssen die Texte in den semantischen Raum projiziert werden, ein Vorgang, der als „Folding in" bezeichnet wird (Martin & Berry, in press). Hierbei werden wiederum auf die Texte die gleichen Filter angewendet wie bei der Berechnung des semantischen Raumes. Schließlich werden die Vektoren der Wörter des Textes (unter Berücksichtigung ihrer Frequenz, der bei der Generierung des semantischen Raumes berechneten Gewichte und der Singulärwerte) addiert. Dabei geht ihre Richtung und Länge in den Gesamtvektor ein.

Die Besonderheit der LSA liegt darin, dass zwei Texte, die von derselben Thematik handeln als ähnlich eingestuft werden, selbst wenn sie keine übereinstimmenden Wörter aufweisen. So korrelieren beispielsweise die inhaltlich identischen Sätze „Pinguine sind am Boden lebende Vögel, die sich von Fisch und Krill ernähren" (Satz 1) und „Ein Pinguin ist ein flugunfähiger Vogel, der Fische und Krebse frisst." (Satz 2) mit .763, obwohl sie mit Ausnahme des ohnehin ausgefilterten Wortes „und" keine gemeinsamen Wörter haben. Der erste Satz korreliert dagegen nur zu .563 mit „Wale sind im Meer lebende Säugetiere, die sich von Fisch und Krill ernähren." (Satz 3), obwohl diese beiden Sätze große Überlappungen aufweisen, aber teilweise von verschiedenen Themen handeln. Die immer noch relativ hohe Korrelation von .563 kommt dadurch zu Stande, dass beide Tiergattungen die gleiche Nahrungsquelle haben und somit eine gewisse inhaltliche Nähe gegeben ist. Der inhaltlich nicht verwandte Satz „Elefanten leben in der afrikanischen Steppe und im indischen Dschungel" korreliert dagegen mit Satz 1 nur noch zu .105 (Demonstration siehe Lenhard, Baier, Schneider & Hoffmann, 2006).

Möglichkeiten und Grenzen LSA-basierter Systeme

Die LSA weißt im Vergleich zu menschlichem Sprachverstehen einige grundlegende Einschränkungen auf. Zunächst muss betont werden, dass Wortbedeutungen nur insoweit repräsentiert werden als sie sich in ihrer gemeinsamen Verwendung widerspiegeln. LSA ist gewissermaßen die mathematische Realisierung der Idee, die Bedeutung von Wörtern (und Texten) durch ihren Gebrauch zu definieren (Wittgenstein, 1953). Das Verfahren verzichtet vollständig auf jegliche Bezüge zu realen sensorischen Wahrnehmungen und Erfahrungen (Landauer & Dumais, 1997).

Des Weiteren werden ausschließlich Relationen des Auftretens von Wörtern in Texten repräsentiert und jegliche Syntax und die damit transportierte Information ausgeblendet. Die beiden fiktiven Anweisungen für Auftragskiller „Tina, nicht Thomas muss liquidiert werden!“ und „Tina nicht, Thomas muss liquidiert werden!“ sind für ein LSA-basiertes System völlig identisch. Nach Filterung von Stoppwörtern würden die Sätze „liquidiert Tina Thomas“ lauten, was wohl für beide Betroffenen unangenehme Konsequenzen hätte. Aussagen, die auf logischen Beziehungen fundieren, wie z. B. mathematische Abhandlungen, verlieren sicher hierdurch ihren wesentlichen Informationsgehalt. Gleichermaßen kann die LSA nicht zwischen Unter- und Oberbegriffen unterscheiden und Negationen erfassen. Es ist deshalb ebenfalls unzulässig LSA-basierten Ähnlichkeitsberechnungen mit menschlichen Assoziationsstrukturen gleichzusetzen (Landauer, Foltz & Laham, 1998). Je nach verwendeter Textsammlung ermittelt ein LSA-System z. B. als ähnliche Wörter zu dem Wort „Erdbeben“ Treffer wie „Beben“ (r = .961), „Kontinentalplatte“ (r = .933), „Erdbebenherd“ (r = .924) usw. Zwar sind diese Treffer plausibel, jedoch assoziieren Menschen vermutlich spontan eher Wörter wie „Katastrophe“, „Opfer“ und „Suchmannschaften“. Darüber hinaus ist es sehr unwahrscheinlich, dass menschlichem Spracherwerb die gleichen Algorithmen zu Grunde liegen, die in der LSA angewandt werden.

Trotz dieser Einschränkungen sollten die Möglichkeiten LSA-basierter Systeme nicht unterschätzt werden. Es gibt zahlreiche Anwendungsfelder, bei denen sie erfolgreich in der Simulation menschlicher Sprachverständnisleistungen angewendet werden konnte. Hierzu gehören die automatische Bewertung von Aufsätzen (Landauer, Laham, Rehder & Schreiner, 1997), Bewertung der Kohärenz und Verständlichkeit von Texten (Foltz, Kintsch & Landauer, 1998), Vorhersage des Lernerfolgs beim Lesen eines Textes auf der Basis des Vorwissens des Lesers und Auswahl geeigneten Lernmaterials (Wolfe et al., 1998), erfolgreiches Bestehen von Multiple-Choice-Wissenstests wie z. B. dem „Test of English as a Foreign Language“ (Landauer & Dumais, 1997), und intelligenten Lernsystem, die inhaltliche Rückmeldung über Textzusammenfassungen geben (Wade-Stein & Kintsch, 2004). Die LSA stellt somit zwar keine gültige Simulation semantischen Wissens und verbaler Intelligenz dar, sie ist jedoch eine hinreichend gute Annäherung für eine Reihe von Anwendungsgebieten.

Übertragung der latenten semantischen Analyse auf das Deutsche

Die im Folgenden dargestellten Untersuchungsergebnisse sind im Rahmen eines Forschungsprojektes entstanden, dessen Ziel die Übertragung der bisher vornehmlich im Englischen angewendeten LSA auf das Deutsche ist. Zwar lassen sich die statistischen Methoden der LSA auf Texte in jeder Sprache anwenden. Gleichwohl gibt es im Deutschen strukturelle Unterschiede zum Englischen, die sowohl Vorteile als auch Nachteile für die LSA mit sich bringen. Unserer Meinung nach am bedeutsamsten sind hierbei Kompositabildung und Flexionsbildung. Während zusammengesetzte Wörter oftmals eine sehr spezifische Bedeutung haben, und sich aus diesem Grund auf die LSA günstig auswirken, führen die hohe Anzahl an Flexionen zu einer sehr starken Zunahme des Lexikons. Die einzelnen Flexionen eines Wortes treten zudem seltener auf oder kommen in der Textsammlung u. U. gar nicht vor, sodass bei der Analyse neuen Textmaterials zahlreiche Wörter fehlen können. Die generelle Lemmatisierung aller Wörter, also die Umwandlung von Wörtern in ihre lexikalische Grundform, hat sich demgegenüber jedoch nicht als überlegen erwiesen. Die gezielte Lemmatisierung von Verben scheint jedoch zu einer Leistungssteigerung des Systems zu führen (Denhière & Lemaire, 2006).

Um zu überprüfen, ob eine auf das Deutsche angewandte LSA offene Antworten adäquat beurteilen kann, wird zunächst dessen Leistung in einem Wissenstest im Multiple-Choice-Format überprüft, um anschließend die Übereinstimmung mit menschlichen Bewertern hinsichtlich der Punktevergabe für Klausurfragen und Textzusammenfassungen zu ermitteln.

Untersuchung 1

Da in der LSA Texte als eine Sammlung von Wörtern betrachtet werden, deren Bedeutungen additiv zusammenwirken, soll zunächst untersucht werden, wie gut das Verfahren die Beziehung zwischen einzelnen Wörtern repräsentieren kann. Ziel ist es zu bestimmen, ob auf der Basis einer mittels LSA automatisch verarbeiteten Textsammlung die automatische Zuordnung von Wörtern zu Kategorien möglich ist und wie gut die LSA hierbei im Vergleich zu Menschen abschneidet. Hierzu wurde die Leistung von Schülern und Studenten (Diplom-Biologie) bei der Kategorisierung von Tierarten mit der Leistung eines LSA-basierten Systems verglichen.

Stichprobe

Die Untersuchung fand im Raum Würzburg statt. Es nahmen 249 Schüler der 5. bis 10. Klasse eines Gymnasiums

(81 Jungen, 73 Mädchen) und der 5. bis 8. Klasse einer Hauptschule (53 Jungen, 44 Mädchen) teil.

Die Studierenden wurden im Biologie-Zentrum der Universität Würzburg rekrutiert. Die Studierenden wurden gefragt, ob sie zur Teilnahme an einer kleinen psychologischen Untersuchung bereit wäre und erhielten für die Teilnahme einen Schokoladenriegel. Es nahmen ausschließlich Studierende des Studiengangs Biologie (Diplom) mit deutscher Muttersprache teil (14 Männer, 30 Frauen). Das Alter variierte zwischen 20 und 27 Jahren (m = 22.8, sd = 1.96).

Verfahren

Die Schüler und Studierenden bearbeiteten ein Wissensquiz, das insgesamt 90 Items mit Tierarten umfasste (Aal, Adler, Alligator, …, Zander, Zwergpinguin), die den Tierklassen Amphibium, Fisch, Insekt, Reptil, Säugetier, Spinne und Vogel zugeordnet werden sollten. Der verwendete semantische Raum basierte auf Texten aus den Themengebieten Biologie, Geologie und Geographie mit Texten aus Schulbüchern, Lexika und Internet-Seiten. Die Texte wurden zum größten Teil automatisch in Abschnitte geteilt, Stopp-Wörter gefiltert und alle Schreibungen in Kleinschreibung konvertiert. Alle Wörter, die weniger als drei Mal auftraten wurden eliminiert. Die Frequenzmatrix umfasste 37 773 Dokumente mit 83 369 verschiedenen Wörtern (Größe des gesamten Korpus 2 178 432 Wörter). Nach der Anwendung einer Log-Entropie-Gewichtung wurde eine SVD gerechnet und 1 000 Dimensionen extrahiert. Diese maximale Anzahl an Dimensionen ergab sich aus den Beschränkungen des Arbeitsspeichers des für die Singulärwertzerlegung verwendeten Rechners (PC mit Pentium IV-Prozessor, Taktfrequenz 3.2 GHz, 3 GB RAM). Die Berechnung des semantischen Raumes dauerte 124 min 50 sec.

Durchführung

Die einzelnen Tierklassen wurden den Schülern und Studierenden zunächst erklärt und sie wurden angewiesen die richtige Tierklasse zu markieren und zu raten, falls sie eine Tierart nicht kannten. Das Quiz wurde ohne Zeitbegrenzung durchgeführt und von allen Versuchspersonen spätestens nach 15 min abgeschlossen. Für jedes richtig gelöste Item wurde in der Auswertung ein Punkt vergeben, sodass maximal 90 Punkte erreichbar waren.

Anschließend wurde das Quiz durch das LSA-basierte System bearbeitet. Eine Aufgabe galt als richtig gelöst, wenn die Tierart im semantischen Raum vorhanden war und mit der korrespondierenden Tierklasse die höchste Korrelation aufwies. Zur Ermittlung der optimalen Anzahl an Dimensionen (höchste Trefferquote) wurde die verwendete Anzahl an Dimensionen systematisch variiert (die ersten 10, die ersten 20, die ersten 30, …, alle 1 000 Dimensionen). Die resultierenden 63 000 Einzelwortvergleiche (90 Items * 7 Auswahlalternativen * 100 Durchläufe des Wissensquiz) dauerten .346 sec (verwendeter Computer: Laptop Dell Latitude D810, 1 GB Arbeitsspeicher, Prozessor Pentium M, Taktfrequenz 1.86 GHz).

Ergebnisse

Die Trefferquote des LSA-basierten Systems nimmt zunächst steil zu und erreicht bei 310 Dimensionen eine maximale Trefferquote von 96.7 %. Danach fällt die Trefferquote leicht ab. Die mittlere Trefferquote liegt bei 88.6 % (sd = .080, Abb. 2).

Bei den schulischen Stichproben wird die maximale Trefferquote in der 10. Klasse (Gymnasium), bzw. 6. Klasse (Hauptschule) erreicht. Das LSA-System erzielte eine hochsignifikant höhere Trefferquote als alle schulischen Stichproben (Tabelle 1). Vergleicht man die maximal erreichten Werte, so lag auch hier das LSA-System über den schulischen Leistungen: Während die zwei besten Schüler eine Trefferquote von 93.3 % erreichten, lag das LSA-System bei einer Dimensionszahl von 310 mit 96.7 % darüber. Die Studierenden lösten im Schnitt 2.3 % mehr Aufgaben als das LSA-System. Der Unterschied zwischen LSA und Studierenden war jedoch nicht signifikant. Insgesamt 5 der 44 Studierenden erzielten eine höhere Punktzahl als die LSA.

Diskussion

Das LSA-System war schulischen Stichproben deutlich überlegen und zeigte einen hohen Grad an Expertise. Die

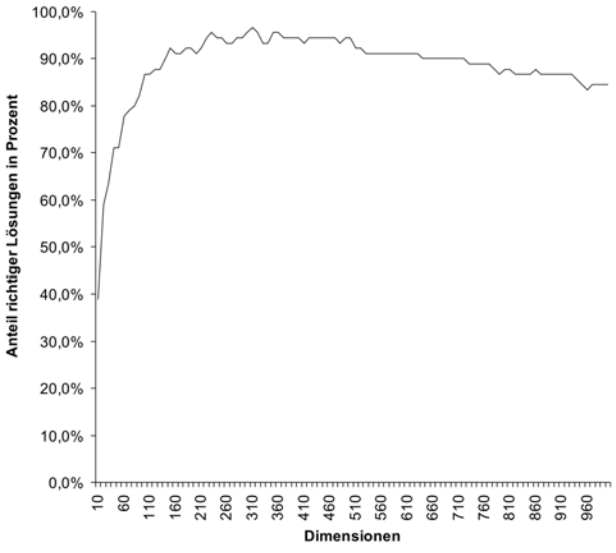

Abbildung 2. Diagramm der Trefferquote des LSA-basierten Systems in einem Tierartenwissenstest. Die Anzahl der verwendeten Dimensionen wurde systematisch in 10er Schritten variiert (1 bis 10, 1 bis 20, …, 1 bis 1000). Die Trefferquote nimmt zunächst steil zu und erreicht bei 310 Dimensionen eine maximale Trefferquote von 96.7 %. Danach fällt die Trefferquote wieder leicht ab.

Tabelle 1. Vergleich der Leistung eines LSA-basierten Systems mit der Leistung von Schülern verschiedener Altersstufen und Schulformen, sowie Studierenden des Studiengangs Biologie (Diplom)

Gruppe	Klasse	*m*	*sd*	*N*	*t*
Hauptschule	5	52	9.45	23	15.71 (*df* = 121)***
	6	57.5	12.13	24	11.86 (*df* = 123)***
	7	49.1	12.58	35	17.00 (*df* = 134)***
	8	53.7	8.64	13	12.03 (*df* = 111)***
Gymnasium	5	53.5	6.86	24	16.20 (*df* = 122)***
	6	60.2	11.86	29	11.02 (*df* = 127)***
	7	58.5	11.60	24	11.44 (*df* = 122)***
	8	62.9	9.39	27	10.11 (*df* = 125)***
	9	62.4	9.06	32	11.16 (*df* = 135)***
	10	66.0	22.51	15	4.84 (*df* = 114)***
Biologie (Diplom)		81.9	5.66	44	1.74 (*df* = 142)

Anmerkungen: Ergebnisse eines Tierartenwissenstest, bei dem Tierarten der jeweiligen Tierklasse zugeordnet werden müssen. Insgesamt sind 90 Punkte erreichbar. Das LSA-System erzielte bei einer systematischen Variation der verwendeten Dimensionen (1 bis 10, 1 bis 20, …, 1 bis 1000) einen durchschnittlichen Wert von $m = 79.8$ ($sd = 7.19$) und schnitt damit hochsignifikant besser ab als alle schulischen Vergleichsgruppen. Studierende (Diplom-Biologie) schnitten nicht signifikant besser ab als das LSA-System. *** signifikant auf einem Niveau von $p = .001$.

LSA erreichte im Schnitt Werte, die an der oberen Leistungsgrenze von Schülern der gymnasialen Oberstufe lagen. Sie schnitt beim Klassifizieren von Tierarten auf einem Niveau ab, das dem Wissensstand von Studierenden der Biologie (Diplom) entspricht.

Zwar könnte eine solche Aufgabenstellung bei wesentlich geringerem Aufwand auch durch ein regelbasiertes System bewältigt werden, bei dem die korrekten Antworten fest kodiert wurden. Die Bedeutung des Ergebnisses liegt demgegenüber insbesondere darin, dass das in den Texten gespeicherte Wissen durch die LSA vollautomatisch extrahiert wurde und das Programm auch bei anderen Aufgaben des gleichen Wissensgebietes vergleichbare Leistungen zeigen kann. Ein regelbasiertes System ist dagegen nicht in der Lage Aufgaben zu lösen, die von den vorab spezifizierten, fest eingegebenen Regeln abweichen. Die LSA ist also für Kategorisierungsaufgaben z. B. im Rahmen von Multiple-Choice-Tests sehr gut geeignet und kann folglich die inhaltliche Beziehung einzelner Wörter gut abbilden. Das Optimum wird bei 310 Dimensionen erreicht. Die Trefferquote bleibt auch bei höheren Dimensionen weitgehend stabil.

Untersuchung 2

Ziel der Untersuchung war die Überprüfung des Zusammenhangs zwischen der Bewertung von offenen Antworten in Prüfungsklausuren durch Menschen und durch ein LSA-basiertes System.

Stichprobe

Es lagen die Antworten von 40 Studenten und Studentinnen aus einer mehrere Jahre zurückliegenden Vordiplomsklausur des Themengebiets „Allgemeine Psychologie I“ vor. Durchschnittlich erreichten die StudentInnen eine Gesamtnote von 2.2 ($s = 1.05$). In einigen Fragen traten deutliche Deckeneffekte auf.

Verfahren

Der für die Ähnlichkeitsberechnungen verwendete semantische Raum wurde auf der Basis der Texte aus 14 Lehrbüchern der Psychologie (Einführung Psychologie, Allgemeine Psychologie I und Kognitive Psychologie) berechnet. Die Texte der Lehrbücher wurden durch eine studentische Hilfskraft in einzelne Abschnitte aufgeteilt. Es wurden Stopp-Wörter gefiltert, sowie Wörter entfernt, die nur ein einziges Mal auftraten. Im semantischen Raum verblieben 66 611 verschiedene Wörter in 27 688 Dokumenten (insgesamt 1 316 599 Wörter). Die Berechnung eines semantischen Raumes mit 1 000 Dimensionen dauerte 80 min 35 sec (PC mit Pentium IV-Prozessor, 3.2 GHz, 3 GB RAM).

Die Klausur umfasste 20 Aufgaben, die im Essay-Stil beantwortet wurden. Für die Beantwortung der Fragen hatten die Prüfungskandidaten insgesamt 120 Minuten Zeit. Die Klausuren wurden anschließend von Experten bewertet, die für den Inhalt der Aufgaben anhand eines vordefinierten Schemas Punkte vergaben. Die maximal erreichbare Punktzahl variierte von Aufgabe zu Aufgabe je nach Schwierigkeitsgrad von zwei bis acht Punkten. Von den 20 Fragen wurden die studentischen Lösungen jener acht Aufgaben am Computer eingegeben, in denen sechs Punkte (sieben Aufgaben) oder acht Punkte (eine Aufgabe) erreicht werden konnten (Aufgaben siehe Tab. 2). In den anderen zwölf Aufgaben konnten zwei bis vier Punkte erreicht werden. Diese Aufgaben wurden auf Grund ihres eingeschränkten Wertebereiches nicht verwendet. Die

Klausuren wurden anonymisiert, Rechtschreibfehler korrigiert und Abkürzungen durch die volle Schreibweise ersetzt (z. B. „Serial reaction time“ statt „SRT“ und „Versuchspersonen“ statt „VPn“), sowie schematische Zeichnungen entfernt.

Als Vergleichstext für die Fragenbeantwortung diente der zugehörige Text des von Prof. Hoffmann erstellten Vorlesungsskripts, das für die Studierenden ein wichtiges Element der Prüfungsvorbereitung darstellte. Der jeweilige Abschnitt des Skripts, auf den sich eine Frage der Vordiplomsklausur bezog, wurde als Vergleichstext für die automatisierte Bewertung der Vordiplomsklausuren herangezogen.

Durchführung

Die Ähnlichkeitsberechnungen wurden auf einem DELL Latitude D810 (1 GB RAM, Intel Pentium M 1,86 GHz) durchgeführt. Es wurde der Kosinus zwischen der studentischen Antwort und dem zugehörigen Text des Vorlesungsskripts berechnet. Fehlende Wörter wurden automatisch lemmatisiert (Algorithmus nach Caumanns, 1999 und der Implementierung in Lucene Apache 2.0, 2006) und die Grundform des Wortes verwendet, falls diese im semantischen Raum enthalten war. Nicht bearbeitete Fragen (Leerantworten) erhielten 0 als Bewertung, da ein Kosinus von 0 lineare Unabhängigkeit bedeutet. Die Anzahl der verwendeten Dimensionen wurde wieder in 10er-Schritten

Tabelle 2. Korrelation zwischen erreichter Punktzahl in einer Klausur und Bewertung durch LSA

Aufgabe	*m*	*sd*	$r_{\text{LSA-Punkte}}$ a)	$r_{\text{LSA-Punkte}}$ b)
5. Lernen auch Menschen latent (unbeabsichtigt)? Begründen Sie Ihre Antwort und diskutieren Sie die Rolle der Aufmerksamkeit beim latenten Lernen. (6 Punkte)	4.3	1.50	.512 (*N* = 40)**	.604 (*N* = 41)**
7. Auf Grund welcher Beobachtungen werden zwei Phasen von schnellen Zielbewegungen unterschieden? Diskutieren Sie die Bedeutung visuellen Feedbacks für die beiden Phasen. (6 Punkte)	4.9	.97	.742 (*N* = 40)**	.812 (*N* = 41)**
8. Wie verändert sich die Verhaltenskontrolle, wenn propriozeptives Feedback unterbunden wird (Deafferentation)? Welche Schlussfolgerungen lassen sich ziehen? (6 Punkte)	4.8	1.33	.670 (*N* = 39)**	.760 (*N* = 41)**
9. Vorinformationen über auszuführende Handlungen führen zu Latenzzeitverkürzungen. Beschreiben Sie ein typisches Experiment (unabhängige und abhängige Variablen, Ergebnisse, Interpretation) zum Vorinformationsparadigma. (6 Punkte)	5.2	1.14	.484 (*N* = 39)**	.874 (*N* = 41)**
11. Vergleichen Sie das Reafferenzprinzip mit der Feedforward-Kontrolle einer Zielbewegung. (8 Punkte)	4.8	1.97	.766 (*N* = 38)**	.816 (*N* = 41)**
12. Charakterisieren Sie die beiden bestehenden Theorien der Farbwahrnehmung und erläutern Sie auf welche psychologischen und physiologischen Beobachtungen sich die beiden Theorien jeweils stützen können. (6 Punkte)	5.4	.75	.237 (*N* = 40)	.764 (*N* = 41)**
16. Begriffe sind mentale Repräsentationen für Klassen von Objekten (Reizwirkungen). Wodurch bestimmt sich, für welche Objekte eine einheitliche Repräsentation ausgebildet wird? Welche Merkmale dominieren in begrifflichen Repräsentationen? (6 Punkte)	4.6	1.50	.374 (*N* = 39)*	.690 (*N* = 41) **
20. Gibt es Gedächtnis (Lernen) ohne Erinnerung? Begründen Sie ihre Antwort und diskutieren Sie Schlussfolgerungen für die Unterscheidung von Gedächtnissystemen. (6 Punkte)	3.9	1.89	.499 (*N* = 36)*	.669 (*N* = 41)**
Summenscore	37.7	6.58	.729 (*N* = 36)**	.804 (*N* = 40)**

Anmerkungen: Es wurde die Ähnlichkeit zwischen studentischen Antworten in acht Klausurfragen einer Vordiplomsklausur Allgemeine Psychologie I und den korrespondierenden Lehrbuchtexten berechnet (semantischer Raum mit 1000 Dimensionen). a) bei Entfernung aller Leerantworten. b) bei Beibehaltung aller Leerantworten. ** signifikant auf einem Niveau von .01, * signifikant auf einem Niveau von .05.

variiert (1 bis 10, 1 bis 20, …, 1 bis 1 000). Dies resultierte in 32 000 Textbewertungen (40 Personen * 8 Aufgaben * 100 Berechnungen). Die Gesamtdauer des Einlesens der Antworttexte, ihrer Projektion in den semantischen Raum und der Ähnlichkeitsberechnungen betrug insgesamt 3.96 sec.

Anschließend wurde die Korrelation zwischen den Bewertungen durch die LSA und den real erzielten Punkten in der Klausur einmal unter Ignorierung der Leerantworten und einmal unter Beibehaltung berechnet. Da bei drei Aufgaben keine Leerantworten auftraten, wurde zur besseren Vergleichbarkeit ein fiktiver Fall hinzugefügt, bei dem alle Antworten leer waren, und der somit eine Art Untergrenze in der Punktevergabe und LSA-Bewertung setzt. In den Summenscores wurde dieser Fall jedoch nicht berücksichtigt um Scheinkorrelationen zu vermeiden.

Ergebnisse

Die Korrelation zwischen menschlicher Bewertung und LSA nimmt zunächst in Abhängigkeit der Anzahl der Dimensionen stark zu und stabilisiert sich ab 300 Dimensionen. Zwischen 300 und 1000 Dimensionen gibt es nur kleine Zu- und Abnahmen, weswegen im Folgenden aus Gründen der Einheitlichkeit die Korrelationen bei der Verwendung von 1 000 Dimensionen als Referenz angegeben werden.

Bei Beibehaltung der Leerantworten ergeben sich ausschließlich hohe bis sehr hohe Korrelationen, zwischen .604 bis .874. Die Zusammenhänge bewegen sich bei Ausschluss von Leerantworten vom unteren bis oberen Bereich (vgl. Tab. 2). Korrelationen unter .4 traten bei Frage 12 und Frage 16 auf. Dies ist auf Deckeneffekte in der menschlichen Bewertung dieser Fragen zurückzuführen: Bei Frage 12 hatten 87.8 % der Studenten eine Punktzahl von 5 oder 6 erreicht, bei Frage 16 waren es 58.5 %.

Die Korrelation der Summenscores weisen in beiden Fällen mit .729 und .804 sehr gute Übereinstimmungen auf. Legt man die Empfehlung von Lienert und Ratz (1998, S. 269) zu Grunde, nach der für die Beurteilung individueller Differenzen Reliabilitätskennwerte von mindestens .7 erforderlich sind, dann erreicht eine LSA-basierte Bewertung bei der Bildung von Summenscores die Verlässlichkeit standardisierter Verfahren.

Diskussion

Die automatische LSA-basierte Bewertung von offenen Antworten in Klausuren erbrachte eine mittlere bis hohe Übereinstimmung mit menschlicher Punktevergabe. Problematisch waren lediglich zwei Aufgaben, bei denen fast alle StudentInnen eine hohe Punktzahl erreichten so dass die interindividuelle Variation für einen aussagefähigen Vergleich mit den Bewertungen durch die LSA nicht ausreichte. Das Problem tritt nicht auf, wenn durch die Berücksichtigung von Leerantworten die untere Grenze des Wertebereichs gesetzt wird. Diese Herangehensweise ist zulässig, da leere Antworten in Klausuren auch real mit 0 Punkten bewertet werden. Tatsächlich wäre es für eine automatische Bewertung notwendig, ebenfalls die obere Grenze des Wertebereichs, z. B. durch eine Musterlösung vorzugeben. Schließlich handelt es sich bei den dargestellten Ergebnissen lediglich um Korrelationen, die für sich noch keine Interpretation der absoluten Höhe der automatischen Bewertung zulassen. Angesichts der Tatsache, dass die teilweise vorhandenen schematischen Zeichnungen durch die LSA nicht erfasst werden konnten während sie bei der Punktevergabe selbstverständlich berücksichtigt wurden, stellt eine Übereinstimmung der Summenscores von .804 ein sehr gutes Ergebnis dar, das in dieser Höhe von Übereinstimmungen zwischen menschlichen Bewertern nur schwer übertroffen werden kann (siehe auch Untersuchung 3).

Untersuchung 3

Im Rahmen einer Untersuchung zur Entwicklung eines computerbasierten Aufsatzassistenten schrieben Studenten Zusammenfassungen zu vorgegebenen Sachtexten, die von menschlichen Bewertern und einem LSA-basierten System beurteilt wurden.

Stichprobe

An der Untersuchung nahmen 51 Studenten und Studentinnen des Studiengangs Psychologie der Universität Würzburg teil. Die Probanden erhielten für ihre Teilnahme Versuchspersonenstunden. Es wurden die Daten von vier Personen ausgeschlossen, deren Muttersprache nicht Deutsch war, sowie von 5 weiteren Personen, die die Instruktion nicht verstanden hatten oder in der vorgegebenen Zeit nicht fertig wurden. Es verblieben insgesamt 42 Personen in der Stichprobe (15 Männer, 27 Frauen). Das Alter variierte zwischen 18 und 27 Jahren (m = 20.6, sd = 1.79).

Verfahren

Die Probanden erarbeiteten am Computer Zusammenfassungen von jeweils zwei Sachtexten. Beide Texte setzten sich aus drei Abschnitten zusammen und bestanden aus 4163 („Kelpwald“) bzw. 3629 Zeichen („Meeresschildkröten“). Die Texte waren in eigenen Worten zusammenzufassen. Die Länge musste zwischen 10 % und 20 % des Originaltextes liegen, wodurch Leerantworten ausgeschlossen wurden. Textstellen, bei denen mehr als vier Wörter in Folge im Originaltext vorkamen, mussten durch die Versuchspersonen geändert werden („Plagiat-Check“).

Für die LSA-basierte Bewertung stand der semantische Raum aus Untersuchung 1 zur Verfügung, also ein Raum der allgemein Texte aus den Bereichen Geologie und Biologie umfasst und somit thematisch die in dieser Untersuchung verwendeten Texte inhaltlich abdeckt ohne spezifisch für sie konstruiert und angepasst worden zu sein.

Für die menschliche Bewertung wurden eindeutige Bewertungsschemata erstellt. Bei der inhaltlichen Bewertung einer Textzusammenfassung konnten bis zu fünf Punkte für die inhaltliche Abdeckung jedes einzelnen Abschnittes des Originaltextes vergeben werden. Die Bewertungen für jeden einzelnen Abschnitt wurden aufsummiert, sodass insgesamt maximal 15 Punkte erreichbar waren.

Durchführung

Die Probanden hatten pro Text 30 Minuten Zeit. Bei einem der beiden Texte erhielten die Probanden abschnittsweise Rückmeldung über die Inhaltsabdeckung (Korrelation mit einer Musterlösung bei einer Dimensionszahl von 350). Im Anschluss an das Experiment wurden die VPn auf einer dreistufigen Skala (nein, unentschieden, ja) befragt, ob ihrer Meinung nach die Rückmeldung des Programms die inhaltliche Güte der Zusammenfassung widerspiegelt. Anschließend wurden die Zusammenfassungen ausgedruckt und in randomisierter Reihenfolge von drei geschulten Bewertern beurteilt, sowie die Korrelationen mit der LSA-Bewertung berechnet. Die Dimensionszahl wurde dabei wieder systematisch variiert (insgesamt 10 200 Berechnungen, vgl. Untersuchung 2, Dauer der Berechnung 1.502 sec, DELL Latitude D810, Pentium M 1.86 GHz, 1 GB RAM). Für die Bewertung der Güte der 102 Zusammenfassungen benötigte jeder Bewerter zwischen 5.0 und 6.5 Stunden.

Ergebnisse

Die Länge der Zusammenfassungen betrug $m = 770.4$ ($sd = 103.9$, Text 1) bzw. $m = 685.1$ ($sd = 90.4$, Text 2) Zeichen. Während die mittleren Interraterkorrelation im Schnitt .688 (Text 1) und .816 (Text 2) betrug, belief sich die durchschnittliche Korrelation zwischen der LSA-basierten Bewertung und der Punktevergabe durch die Bewerter auf .629 (Text 1) und .640 (Text 2, siehe Tab. 3).

Menschliche Bewerter erreichten im Mittel weder bei Text 1 noch bei Text 2 eine höhere Interraterkorrelation als zwischen LSA und menschlichem Bewerter.

Befragt nach ihrer Meinung zum inhaltlichen Feedback gaben 69.05 % der Probanden an, dass das inhaltliche Feedback für das Erstellen der Zusammenfassung sehr hilfreich gewesen sei, jedoch waren nur 23.81 % der Meinung, dass es die Güte der Zusammenfassung angemessen widerspiegeln würde.

Diskussion

Die Bewertungen durch die LSA wiesen eine ähnlich hohe Übereinstimmung mit der Bewertung durch menschliche Bewerter auf wie die Interraterkorrelationen der Menschen untereinander. Eine automatische Bewertung erwies sich folglich als genauso verlässlich wie die Punktevergabe durch Menschen. Trotz dieser hohen Validität waren die Probanden nicht von der Güte der inhaltlichen Rückmeldung überzeugt. Dies weist auf ein grundlegendes Problem automatischer Bewertung hin: Trotz der Eliminierung verzerrender menschlicher Bewertungsheuristiken und Stereotype fällt es Menschen schwer, die automatische inhaltliche Bewertung von Texten zu akzeptieren, auch wenn diese sich als sehr valide erweist. Gleichzeitig zeigte sich, dass auch bei menschlichen Bewertern keine absolute Auswertungsobjektivität erwartet werden kann, selbst wenn es sich um relativ kurze Texte handelt und das Bewertungsschema sehr genau definiert ist.

Allgemeine Diskussion

Es konnte gezeigt werden, dass mittels LSA semantisches Wissen angemessen repräsentiert werden kann. Trotz Einschränkungen hinsichtlich fehlender Syntax konnten bei der automatischen Bewertung von Klausurergebnissen und Textzusammenfassungen gute bis sehr gute Übereinstimmung mit menschlichen Bewertern erzielt werden. Bei der Verwendung von Summenscores werden Werte erreicht, die für standardisierte psychologische Testverfahren erforderlich sind. Die Anwendung der LSA auf das Deutsche ist somit gut möglich. Menschliche Bewertung erwies sich nicht als verlässlicher im Vergleich zu LSA-basierter Bewertung, und das bei einem winzigen Bruchteil der benötigten Zeit. Große Vorteile ergeben sich bei

Tabelle 3. Korrelation zwischen der Bewertung von Textzusammenfassungen durch Menschen und durch LSA

Text	$r_{\text{LSA-A}}$	$r_{\text{LSA-B}}$	$r_{\text{LSA-C}}$	$m_{\text{r(ABC)}}{}^{\text{a, c}}$	$m_{\text{r(LSA)}}{}^{\text{b, c}}$	z^{d}
Kelpwald	.584 **	.687 **	.611 **	.688 **	.629 **	.33
Meeresschildkröten	.540 **	.702 **	.670 **	.816 **	.640 **	1.18

Anmerkungen: Die Zusammenfassungen ($N = 42$) konnten mit bis zu 15 Punkten bewertet werden. $r_{\text{LSA-A,}}$ $r_{\text{LSA-B}}$ und $r_{\text{LSA-C}}$ gibt die Korrelation zwischen der LSA-Bewertung und den einzelnen menschlichen Bewertern (A, B und C) wieder, $m_{\text{r(ABC)}}$ die mittlere Korrelation der Bewerter untereinander und $m_{\text{r(LSA)}}$ die durchschnittliche Korrelation zwischen der LSA-Bewertung und den menschlichen Bewertern. Menschliche Bewerter erzielten keine höhere Übereinstimmung als Menschen verglichen mit LSA. a) durchschnittliche Korrelationen der drei menschlichen Bewerter, b) durchschnittliche Korrelation zwischen menschlichen Bewertern und LSA, c) Mittelwerte von Korrelationen wurden mittels Fisher's Z-Transformation gebildet, d) Test auf Korrelationsunterschiede zwischen $m_{\text{r(ABC)}}$ und $m_{\text{r(LSA)}}$ mit Prüfgröße z (Olkin & Siotani, 1964 nach Bortz, 1999, S. 213). ** signifikant auf einem Niveau von .01, * signifikant auf einem Niveau von .05.

automatischen Textbewertungen folglich, wenn eine große Anzahl elektronisch vorliegender Textantworten in einer kurzen Zeit beurteilt werden müssen, und es in den Texten v. a. um die Überprüfung von Wissensinhalten geht.

Die Nachbefragung der Probanden zeigte, dass die computerbasierte Einschätzung von Aufsätzen zwar als hilfreich angesehen wird, jedoch das Vertrauen gegenüber dieser Einschätzung eher gering ausfällt. Trotz der Tatsache, dass ein Bias hinsichtlich Geschlecht, Text- und Satzlänge etc. ausgeschlossen werden kann, und die Übereinstimmung mit menschlichen Bewertern hoch ist, hat eine automatisierte Aufsatzbewertung folglich möglicherweise ein Akzeptanzproblem bei den Verfassern der Aufsätze. Zukünftige Forschungen werden sich deshalb unter anderem mit der Frage befassen müssen, wie die Rückmeldung zu gestalten ist, damit die Bewertung nachvollziehbar wird. Hierzu könnten beispielsweise die Markierung von irrelevanten Sätzen und redundanten Textstellen zählen, wie dies in der interaktiven LSA-basierten Lernumgebung Summary Street® der Fall ist (Kintsch, Steinhart, Stahl & LSA research group, 2000), oder auch stilistische Informationen (Ishioka & Kameda, 2006).

Neben diesen Forschungsfragen, die auf die Interaktion zwischen Computer und Mensch abzielen, gibt es zahlreiche ungeklärte technische Aspekte. Trotz der eleganten Einfachheit der LSA, die völlig ohne die Spezifikation von Regeln auskommt, gibt es viele Parameter, die ihre Effektivität deutlich beeinflussen. Optimale Werte variieren je nach Aufgabenstellung und lassen sich bislang im Wesentlichen nur post hoc durch Versuch und Irrtum herausfinden. Zu diesen Parametern gehören die Länge der Dokumente in der Textsammlung, die Größe der Textsammlung, die Wahl der Gewichtungsfunktionen, die Aufspaltung der Komposita, die komplette Lemmatisierung der Wörter oder zumindest bestimmter Wortarten, die Eignung von Textarten. Weiterhin stellt sich die Frage, ob themenspezifische Textsammlungen und in Folge auch themenspezifische semantische Räume verwendet werden sollten, oder ob allgemeine Räume ebenfalls hinreichend gute Resultate erbringen. Welche Anzahl an Dimensionen ist optimal und lässt sich in Analogie zum Scree-Test bei der Faktorenanalyse auch a priori ein Kriterium festlegen? Während für einige dieser Fragestellungen, wie z. B. der Wahl der Gewichtungsfunktionen bereits in verschiedenen Sprachen systematische Untersuchungen durchgeführt wurden, kristallisieren sich für andere Erfahrungswerte heraus. Erschwert wird die Situation, da die verschiedenen Variationsmöglichkeiten interdependet sind und somit keine optimalen Einstellungsmöglichkeit einzelner Parameter, sondern meist nur günstige Parameterkombinationen existieren. Es bleibt abzuwarten, ob hier eine theoretische Fundierung möglich ist.

Trotz dieser Schwierigkeiten können nach unserer Einschätzung LSA-basierte Systeme eine wertvolle Hilfe bei der automatischen Bewertung von Aufsätzen leisten, auch wenn v. a. bei kritischen Aufgabenstellungen wie Studieneingangstests oder Diplomnoten die endgültige Entscheidung letztlich beim Menschen verbleiben muss.

Literatur

Apache Lucene 2.0 (2006). Lucene Java. *The Apache Software Foundation.* verfügbar unter: http://lucene.apache.org/ [25.07.2006].

Berry, M. W, Dumais, S. T. & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37* (4), 573–595.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5. vollständig überarbeitete und aktualisierte Auflage). Berlin: Springer.

Bühner, M. (2004). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson.

Burstein, J., Kukich, K., Wolff, S., Lu, C. & Chodorow, M. (1998). Enriching automated scoring using discourse marking. *Proceedings of the Workshop on Discourse Relations and Discourse Marking, 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics.*

Caumanns, J. (1999). A Fast and Simple Stemming Algorithm. *Technical Report Nr. TR-B-99-16 des Fachbereichs Informatik der Freien Universität Berlin*. verfügbar unter: http://www.inf.fu-berlin.de/inst/pubs/tr-b-99-16.abstract.html [25.07.2006].

Chase, C. I. (1979). The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement, 16*, 293–297.

Chase, C. I. (1986). Essay test scoring : interaction of relevant variables. *Journal of Educational Measurement, 23*, 33–41.

Coffman, W. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 5*, 24–36.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). *Indexing by Latent Semantic Analysis. Journal of the American Society For Information Science, 41*, 391–407.

Denhière, G. & Lemaire, B. (2006). *Representing children's semantic knowledge from a multisource corpus.* Vortrag im Rahmen eines Workshops in Würzburg, 05.02.2006.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual Coherence with latent Semantic Analysis. *Discourse Processes, 25*, 285–307.

Haladyna, T. (1999). *Developing and validating multiple-choice test Items*. Mawhwah, NJ: Erlbaum.

Hughes, D. C., Keeling B. & Tuck, B. F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational and Psychological Measurement, 43*, 1047–1050.

Ishioka, T. & Kameda, M. (2006). Automated Japanese Essay Scoring System based on Articles Written by Experts. *Coling-ACL 2006 Conference.*

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C. & Lamb, R. (2000). Developing summarization skills through the use of LSA-Based feedback. *Interactive Learning Environments, 8*, 87–109.

Kontostathis, A. & Pottenger, W. M. (2002). Detecting patterns in the LSI term-term matrix. *Workshop on the Foundation of Data Mining and Discovery, IEEE International Conference on Data Mining.*

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., Rehder, B. & Schreiner, M. E., (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Pro-*

ceedings of the 19th annual meeting of the Cognitive Science Society (pp. 412–417). Mawhwah, NJ: Erlbaum.
Lemaire, B. & Denhière, G. (2004) Incremental Construction of an Associative Network from a Corpus. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings 26th Annual Meeting of the Cognitive Science Society* (pp. 825–830), Chicago.
Lenhard, W., Baier, H. Schneider, W. & Hoffmann, J. (2006). Forschungsprojekt „Förderung des Textverständnisses“: LSA-Modul. verfügbar unter: http://www.summa.psychologie.uni-wuerzburg.de/summa/coa/login/ [20.07.2006].
Lienert, G. & Ratz, U. (1998). *Testaufbau und Testanalyse* (6. Auflage). Weinheim: Beltz.
Marshall, J. C. & Powers, J. M. (1969). Writing neatness, composition errors and essay grades. *Journal of Educational Measurement, 6*, 97–101.
Martin, D. & Berry, M. (in press). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), *The handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
Meyer, G. (1939). The choice of questions on essay examinations. *Journal of Educational Psychology, 30*, 161–171.
Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research, 29*, 495–512.
Nakov, P., Popova, A. & Mateev P. (2001). Weight functions impact on LSA performance. *Proceedings of the EuroConference Recent Advances in Natural Language Processing, RANLP'01*, 187–193.
Olkin, J. & Siotani, M. (1964). Asymptotic distribution functions of a correlation matrix. *CA: Stanford University Laboratory for Quantitive Research in Education. Report No. 6*.
Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappa, 47*, 238–243.
Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K. & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*, 337–354.
Wade-Stein, D. & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333–362.
Wild, F., Stahl, Ch., Stermsek, G. & Neumann, G. (2005). Parameters Driving Effectiveness of Automated Essay Scoring with LSA. *Proceedings of the 9th International Computer Assisted Assessment Conference*, 485–494.
Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.
Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes, 25*, 309–336.

Dr. Wolfgang Lenhard
Prof. Dr. Wolfgang Schneider

Lehrstuhl Psychologie IV
Universität Würzburg
Röntgenring 10
97070 Würzburg
E-Mail: wolfgang.lenhard@mail.uni-wuerzburg.de

Dr. Herbert Baier
Prof. Dr. Joachim Hoffmann

Lehrstuhl Psychologie III
Universität Würzburg
Röntgenring 11
97070 Würzburg