Bewertung von kurzen Freitextantworten in automatischen Prüfungssystemen

Martin Filipczyk, Michael Striewe, Michael Goedicke

Paluno - The Ruhr Institute for Software Technology
Universität Duisburg-Essen, Campus Essen
Gerlingstraße 16, 45127 Essen
vorname.nachname@s3.uni-due.de

Abstract: Die manuelle Bewertung von offenen Aufgaben, insbesondere Freitextaufgaben, ist für Lehrende zeitaufwändig und unterliegt Schwankungen durch subjektive Einschätzungen. Besonders im Hinblick auf aktuelle Entwicklungen wie Massive Open Online Courses (MOOC), bei denen sich zehntausende Studierende für Online-Vorlesungen einschreiben, ist formatives Assessment ohne Automatisierung der Bewertung unmöglich. Der vorgestellte Ansatz kombiniert verschiedene Verfahren zur automatischen Bewertung von Freitextantworten und integriert diese in das Prüfungssystem JACK. Die Lösung bietet dem Lehrenden die Möglichkeit, auch ohne große Mengen von Trainingsdaten Antworten automatisiert bewerten zu lassen und konnte sich in einer ersten, auf echten Klausurdaten basierenden Evaluation bewähren.

1 Einleitung

Eine wichtige Aufgabe für Dozenten ist die Korrektur eingereichter studentischer Arbeiten und Lösungen, welche für offene Fragestellungen zeitintensiv ist und die Gefahr subjektiver Bewertungen birgt [Gru10]. Durch aktuelle Entwicklungen wie *Massive Open Online Courses* ist es bedingt durch die Vielzahl an Teilnehmern unmöglich, sämtliche Einreichungen mit den verfügbaren Ressourcen manuell zu korrigieren. Zusätzlich nehmen Studierende erfahrungsgemäß die Möglichkeit wahr, Lösungen zu Tageszeiten einzureichen, zu denen kein Dozent verfügbar ist [SG11]. Deshalb sind computergestützte Prüfungssysteme heutzutage an der Hochschule weit verbreitet und haben das Potential, die angesprochenen Probleme der manuellen Bewertung studentischer Einreichungen zu lösen. Eine angemessene Bereitstellung technischer Ressourcen vorausgesetzt sind automatische Bewertungen in der Regel innerhalb weniger Sekunden verfügbar. Der Einsatz automatischer Systeme mit algorithmisch festgelegten Verfahren ermöglicht die objektive, parallele und tageszeitunabhängige Bewertung großer Mengen von Einreichungen. Aktuell sind Multiple-Choice-Verfahren in E-Assessment-Systemen weit verbreitet. Das Einbringen von individuellen Lösungswegen, Interpretationen, Meinungen und Begründungen ist dabei für die Studierenden nicht möglich. An dieser Stelle können Freitextaufgaben eingesetzt werden, um eine größere Bandbreite an Fähigkeiten der Studierenden abzurufen und somit eine umfassendere Prüfung zu gewährleisten. Die didaktische Zielsetzung hinter ei-

ner Freitextaufgabe kann sehr unterschiedlich sein und bspw. die Abfrage von inhaltlichem Wissen, die kreative Beschreibung von Zusammenhängen in eigenen Worten oder die Prüfung der grammatischen und orthographischen Fertigkeiten eines Studierenden umfassen.

1.1 Verwandte Arbeiten

e-rater® [BKW+98] ist ein System zur automatischen Bewertung der sprachlichen Qualität längerer Aufsätze und wird u.a. zur automatisierten Auswertung von Aufgaben des TOEFL® (*Test Of English as a Foreign Language*) eingesetzt. Intelligent Essay Assessor [FLL99] bewertet Aufsätze mittels Latent Semantic Analysis (vgl. dazu [LFL98]), benötigt jedoch eine große Menge von Trainingsdaten. c-rater [SB09] bewertet Freitextantworten auf der Grundlage von durch Dozenten erstellten Modellantworten und bedient sich komplexer linguistischer Analysetechniken, weist jedoch Schwächen in der Bewertung kurzer und prägnanter sowie aus mehreren Sätzen bestehenden Antworten auf.

2 Ansatz

Der hier präsentierte Ansatz zur automatischen Bewertung von Freitextantworten stellt eine konfigurierbare Kombination dreier Verfahren, des Bayes-, BLEU- und des RegEx-Verfahrens, dar. Jedes dieser Verfahren liefert zu einer eingereichten Antwort eine Bewertung in Form einer Punktzahl, welche über eine durch den Lehrenden festgelegte Gewichtung zu einer Gesamtpunktzahl kombiniert wird. Diese Verfahren wurden entsprechend der Architektur des Prüfungssystems JACK (vgl. dazu [SBG09]) implementiert. JACK ist ein am Lehrstuhl „Spezifikation von Softwaresystemen“ der Universität Duisburg-Essen entwickeltes System zur Unterstützung summativen und formativen Assessments [SBG09]. Bedingt durch die vielfältigen Anforderungen an die Bewertung von Lösungen verschiedenster Aufgabentypen können voneinander unabhängige Checker-Komponenten entwickelt werden, welche zur Laufzeit in JACK eingebunden werden. Jedes der drei in den vorigen Abschnitten vorgestellten Verfahren wurde in einer solchen Checker-Komponente umgesetzt. Die Bewertung einer Antwort wird aus den durch die Komponenten zurückgegebenen Punktzahlen und einer durch den Lehrenden definierten Gewichtung errechnet.

Das Bayes-Verfahren [RL02] ist eine auf dem Bayestheorem basierende Methode zur automatischen Bewertung von Aufsätzen. Mit Hilfe des Verfahrens können nach einer Trainingsphase Aufsätze in vorher definierte nominalskalierte Bewertungsklassen eingestuft werden. Das Verfahren kann auf verschiedene Typen von Termen wie bspw. Wörter oder n-gramme, das bedeutet aus n Wörtern zusammengesetzte Wortfolgen, angewendet werden. Die Ausgabe des Verfahrens ist eine Abbildung der Bewertungsklassen auf die Wahrscheinlichkeit, mit der sich eine Antwort der jeweiligen Klasse zuordnen lässt. Die Kategorien, nach denen Antworten bewertet werden, bieten dem Lehrenden die Möglichkeit, Antworten nach mehreren Schemata zu klassifizieren. Das Verfahren wurde ausgewählt, da es als statistischer Ansatz auch durch den Lehrenden unvorhergesehene Formulierungen

angemessen einschätzen kann. Ein wesentlicher Nachteil des Verfahrens ist die große Anzahl an Antworten, welche für die Trainingsphase benötigt werden (bspw. 462 in [RL02]) und die für neu entworfene Aufgaben im Normalfall nicht zur Verfügung stehen.

Die *Bilingual Evaluation Understudy* (BLEU) ist in ihrer ursprünglichen Form eine Methode zur automatischen Evaluation von maschinellen Übersetzungen [PRWZ02]. Maschinell übersetzte Sätze eines Textes (sogenannte Kandidaten) werden über n-gramme verschiedener Ordnungen mit mehreren manuell erstellten Referenzübersetzungen derselben Ausgangssätze verglichen. Kandidaten, die in ihrer Länge wesentlich von den Referenzen abweichen, werden durch BLEU abgewertet. Das Problem der automatischen Evaluation von maschinellen Übersetzungen durch die BLEU-Methode kann auf das Problem der automatischen Bewertung von Freitextantworten abgebildet werden, indem studentische Antworten als maschinelle Übersetzungen und von Lehrenden erstellte korrekte Musterantworten als Referenzübersetzung interpretiert werden [PAR04].

Das in den Ansatz integrierte RegEx-Verfahren erweitert die in [BCK+02] beschriebene, auf regulären Ausdrücken basierende Methode. Der Lehrende erstellt für jede Aufgabe eine Bewertungsvorlage, welche eine Menge von Bewertungsschlüsseln enthält. Ein Bewertungsschlüssel besteht aus einem regulären Ausdruck, einer Punktzahl sowie optionalen Alternativen in Form weiterer Schlüssel. Diese Alternativen können mit derselben Punktzahl (bspw. für Synonyme) oder einer niedrigeren Punktzahl (bspw. für ungenauere Aussagen) definiert werden. Zur Bewertung einer Antwort wird der so aufgespannte Baum von Bewertungsschlüsseln durchlaufen und die Punktzahlen derjenigen Schlüssel summiert, deren regulärer Ausdruck für die Antwort greift. Da für das Verfahren keine Trainingsdaten notwendig sind, kann es durch sorgfältige Definition einer Bewertungsvorlage direkt für neue Aufgaben eingesetzt werden. Allerdings werden auch korrekte Antworten, die durch keinen Bewertungsschlüssel abgedeckt werden, automatisch schlecht bewertet.

Für die Auswahl der zu kombinierenden Verfahren wurden weitere Methoden betrachtet. Aufgrund der schlechteren Korrelation mit manuellen Bewertungen [PAR04] wurde das Vector Space Model (vgl. dazu [SWY75]) zugunsten des BLEU-Verfahrens verworfen. Die Latent Semantic Analysis (LSA, vgl. dazu [LFL98]) benötigt ebenso wie das Bayes-Verfahren eine große Menge an Trainingsdaten, das Bayes-Verfahren erhielt aufgrund seiner Konfigurationsmöglichkeiten der Vorzug gegenüber LSA. Das in [PS05] vorgestellte Verfahren wurde aufgrund der hohen Komplexität für den Lehrenden nicht in den Ansatz integriert. Der vorgestellte Ansatz ist nicht auf die drei beschriebenen Verfahren beschränkt, sondern kann zukünftig flexibel um weitere Verfahren erweitert werden.

3 Evaluation

Für die Evaluation wurden zunächst Antworten auf die in Abbildung 1 dargestellte Frage der Veranstaltung „Programmierung“ aus archivierten Klausurunterlagen digitalisiert und dienten dem Training der beschriebenen Verfahren sowie der Kalibrierung, d.h. der Ermittlung geeigneter Gewichtungen der Verfahren zur Optimierung verschiedener Metriken. Mit diesen Gewichtungen wurde der Ansatz auf weitere studentische Lösungen

Abbildung 1: Die für die Evaluation verwendete Aufgabe

Tabelle 1: In der Kalibrierungsphase ermittelte Optimalgewichtungen

Gewichtung	w_{Bayes}	w_{BLEU}	w_{RegEx}
min. Abweichung (min_A)	0,63	0,04	0,33
max. Übereinstimmung (max_U)	0,89	0,00	0,11

angewandt. Die studentischen Antworten wurden unverändert inklusive aller orthografischen und grammatikalischen Fehler übernommen. Die manuellen Bewertungen der digitalisierten Klausurantworten wurden an die Punktevergabe in JACK angepasst, da die Studierenden in der Klausur bis zu 5 Punkten erreichen konnten, während einer Lösung in JACK bis zu 100 Punkte zugewiesen wird (im Folgenden bezieht sich „Punkte“ auf eine in JACK erreichbare Punktzahl, während der Begriff „Punktstufe“ für in der Klausur vergebene Bewertungen verwendet wird). Insgesamt wurden 161 Antworten mit einer mittleren Länge von 19 Wörtern digitalisiert, von denen unter Berücksichtigung der Verteilung der Antworten bzgl. ihrer Punktzahl zufällig 128 Antworten (ca. 80%) für die Trainingsphase und 33 Antworten (ca. 20%) für die Ermittlung der Optimalgewichtungen verwendet wurden. In der Kalibrierungsphase wurden diese 33 Antworten durch die vorher trainierten Verfahren einzeln bewertet. Aus den Bewertungen wurde jeweils eine Gewichtung berechnet, welche das arithmetische Mittel der Abweichungen von manuellen Bewertungen minimierte bzw. die Anzahl der Übereinstimmungen maximierte. Durch Anwendung der Gewichtungen zur Ermittlung von Gesamtpunktzahlen traten Bewertungen außerhalb der Menge $\{0, 20, 40, 60, 80, 100\}$ auf. Um die Vergleichbarkeit mit den manuellen Bewertungen zu gewährleisten, wurden diese Punktzahlen auf die nächstgelegene Punktzahl aus dieser Menge auf- bzw. abgerundet. Die ermittelten optimalen Gewichtungen können Tabelle 1 entnommen werden. Das BLEU-Verfahren weist zumindest in der verwendeten Konfiguration keinen bzw. nur sehr geringen Anteil an den optimalen Gewichtungen auf. Sowohl für die minimale Abweichung min_A als auch für die maximale Anzahl an Übereinstimmungen max_U dominiert das Bayes-Verfahren.

Die Antworten für die Durchführung des Testlaufs wurden gesammelt, indem die Aufgabe aus Abbildung 1 in eine JACK-Instanz eingepflegt und Studierenden zur Klausurvorbereitung angeboten wurde, wobei die Einreichungen manuell korrigiert wurden. Dadurch wurden insgesamt 16 einzigartige Antworten mit durchschnittlich 29 Wörtern gewonnen. Tabelle 2 zeigt die im Testlauf ermittelten Ergebnisse. Die Metriken, nach denen die Berechnungsmodelle bewertet wurden, waren die mittlere Abweichung der automati-

Tabelle 2: Ergebnisse des Testlaufs unter Verwendung verschiedener Gewichtungen

Bewertungsmodell	mittl. Abw.	Stdabw.	Übereinst.	Abw. ≤ 20
Median	18,750	22,472	50,00%	68,75%
min. Abweichung (min_A)	10,000	10,328	50,00%	**100,00%**
max. Übereinst. (max_U)	**5,000**	**8,944**	**75,00%**	**100,00%**

schen von den manuellen Bewertungen, die Standardabweichung dieser Abweichungen, die Quote der exakten Übereinstimmungen sowie die Quote der Abweichungen von maximal 20 Punkten (entspricht einer Punktstufe innerhalb der Klausur). Beide Gewichtungen erreichen für letzteres einen optimalen Wert von 100% und weichen somit bei allen Antworten der Testdaten um maximal eine Punktstufe ab. Die Gewichtung max_U erzielte mit 75% tatsächlich die beste Übereinstimmungsquote der verglichenen Bewertungsmodelle. Des Weiteren ist diese Gewichtung auch in den verbleibenden beiden Kriterien überlegen. Es ist auffällig, dass min_A gegenüber max_U mit 10 Punkten eine doppelt so hohe mittlere Abweichung aufweist. Möglicherweise ist dies auf Schwankungen zurückzuführen, die durch die geringe Anzahl an Testdaten entstanden sein können.

4 Fazit und Ausblick

Der vorgestellte und in das Prüfungssystem JACK integrierte Ansatz kombiniert mit dem Bayes-, dem BLEU- und dem RegEx-Verfahren drei Methoden zur automatischen Bewertung von Freitextantworten. Durch den Einsatz anpassbarer Gewichtungen können die Stärken der einzelnen Verfahren gewinnbringend ausgenutzt werden. Prinzipiell ist der Ansatz nicht auf Trainingsdaten angewiesen, bei Vorhandensein von Trainingsdaten können jedoch Verfahren aktiviert werden, welche auch unbekannte Formulierungen verarbeiten können. Das verbreitete Problem, dass zur Evaluation eines Ansatzes nicht genügend reale Trainingsdaten zur Verfügung stehen, wurde durch die Digitalisierung archivierter Klausurantworten gelöst. Die über Kalibrierungsdaten optimierten Gewichtungen sind dem ebenfalls untersuchten Median klar überlegen und liefern Übereinstimmungsquoten von bis zu 75% bei der vorliegenden Intervallskala der Größe 6, wobei sämtliche Bewertungen um maximal eine Punktstufe von den manuellen abwichen. Dieses Ergebnis zeigt, dass der Ansatz prinzipiell zur automatischen Bewertung von kurzen Freitextantworten geeignet ist. In einer Folgestudie soll untersucht werden, inwiefern die Ergebnisse der einzelnen Verfahren durch veränderte Konfigurationen weiter verbessert werden können. Insbesondere die Leistung des BLEU-Verfahrens, das verglichen mit [PAR04] unerwartet schlecht abschnitt, könnte durch eine gezieltere Auswahl der Referenzantworten verbessert werden. Die Generierung von Feedback zu eingereichten Lösungen soll zukünftig in den Ansatz integriert werden. Momentan ist es ausschließlich dem Lehrenden möglich, statistische Daten (bspw. Bayes-Wahrscheinlichkeiten oder gefundene Bewertungsschlüssel) einzusehen. Die Studierenden könnten von automatisch erzeugtem Feedback profitieren, da es dabei hilft, die Bewertung ihrer Antworten nachzuvollziehen und somit zur Akzeptanz eines Systems zur automatischen Bewertung von Freitextantworten beiträgt.

Literatur

[BCK+02] L. F. Bachman, N. Carr, G. Kamei, M. Kim, M. J. Pan, C. Salvador und Y. Sawaki. A reliable approach to automatic assessment of short answer free responses. In *Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, Seiten 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[BKW+98] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder und M. D. Harris. Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, Seiten 206–210, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[FLL99] P. W. Foltz, D. Laham und T. K. Landauer. The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), Oktober 1999.

[Gru10] S. Gruttmann. *Formatives E-Assessment in der Hochschullehre - computerunterstützte Lernfortschrittskontrollen im Informatikstudium*. Dissertation, Westfälische Wilhelms-Universität Münster, Januar 2010.

[LFL98] T. K. Landauer, P. W. Foltz und D. Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

[PAR04] D. Pérez, E. Alfonseca und P. Rodríguez. Application of the BLEU method for evaluating free-text answers in an e-learning environment. In *Proceedings of the 4th International Language Resources and Evaluation Conference (LREC-2004)*, Seiten 1351–1354, Lissabon, 2004.

[PRWZ02] K. Papineni, S. Roukos, T. Ward und W.-J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Seiten 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[PS05] S. G. Pulman und J. Z. Sukkarieh. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, Seiten 9–16, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[RL02] L. M. Rudner und T. Liang. Automated Essay Scoring Using Bayes' Theorem. *The Journal of Technology, Learning, and Assessment*, 1(2), Juni 2002.

[SB09] J. Z. Sukkarieh und J. Blackmore. c-rater: Automatic Content Scoring for Short Constructed Responses. In *Proceedings of the Twenty-Second International FLAIRS Conference*. AAAI Press, 2009.

[SBG09] M. Striewe, M. Balz und M. Goedicke. A Flexible and Modular Software Architecture for Computer Aided Assessments and Automated Marking. In *Proceedings of the First International Conference on Computer Supported Eductation (CSEDU), 23 - 26 March 2009, Lisboa, Portugal*, Jgg. 2, Seiten 54–61. INSTICC, 2009.

[SG11] M. Striewe und M. Goedicke. Studentische Interaktion mit automatischen Prüfungssystemen. In *DeLFI 2011 - Die 9. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V.*, number 188 in LNI, Seiten 209–220. GI, 2011.

[SWY75] G. Salton, A. Wong und C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, November 1975.