Ulrike Lucke et al. (Hrsg.): Die 14. E-Learning Fachtagung Informatik,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 203

Evaluation automatisierter Ansätze für die Bewertung von Modellierungsaufgaben

Michael Fellmann[1], Peter Fettke[2], Constantin Houy[2], Peter Loos[2], Andreas Oberweis[3], Andreas Schoknecht[3], Michael Striewe[4], Tom Thaler[2]und Meike Ullrich[3]

Abstract: E-Assessments sind immer weiter verbreitet in der Hochschullehre. In einigen Einsatzgebieten birgt die Erstellung von automatisierten Ansätzen für die Bewertung besondere Herausforderungen, so auch bei klassischen Modellierungsaufgaben. In diesem Beitrag wird ein Schema für die Bewertung von Modellen in Form von Ereignisgesteuerten Prozessketten entworfen und zwei automatisierte Bewertungsansätze einer manuellen Bewertung durch Lehrende gegenübergestellt. Als Ergebnis lassen sich vielversprechende Übereinstimmungen feststellen, die Potenziale für die Anwendung automatisierter Bewertungsverfahren von Modellierungsaufgaben aufzeigen.

Keywords: E-Assessment, automatisierte Bewertung, Modelle, Ereignisgesteuerte Prozesskette.

1 Einleitung

In der Praxis wird konzeptuelle Modellierung beispielsweise im Datenbankentwurf oder in der Geschäftsprozessmodellierung genutzt. Aufgrund ihrer Praxisrelevanz ist sie curricularer Bestandteil zahlreicher Studiengänge, insbesondere im Bereich der Wirtschaftsinformatik [Gl08]. Der Bewertung von Modellierungsaufgaben kommt dabei sowohl in der traditionellen wie auch in der digitalen Hochschullehre eine entscheidende Bedeutung zu. In der traditionellen Präsenzlehre ist eine Bewertung von Modellierungsaufgaben zumeist im Rahmen von Übungen oder schriftlichen Klausuren erforderlich. Dabei besteht die Schwierigkeit zum einen darin, ein Bewertungsschema zu erstellen, welches die komplette Bandbreite möglicher Lösungen abdeckt und gleichzeitig faire, intersubjektiv nachvollziehbare und reproduzierbare Ergebnisse liefert. Zum anderen muss aber auch sichergestellt sein, dass ein Bewertungsschema über alle zu bewertenden Lösungen hinweg konsistent und fehlerfrei angewendet wird. Dies manuell zu bewerkstelligen kann mitunter schwierig sein, wenn z. B. die Korrekturarbeit auf mehrere Personen aufgeteilt wird.

Mit zunehmender Digitalisierung der Lehre und dem Einsatz von E-Assessments können Lösungen zukünftig elektronisch eingereicht und anschließend (teil-)automatisiert bewertet werden [Ha16]. Für Lehrende hätte eine derartige IT-gestützte Bewertung neben Zeit- und Kostenreduktion zusätzlich weitere Vorteile: Die digitale Erfassung von Lösungen

[1] Universität Rostock, Albert-Einstein-Str. 22, 18057 Rostock, michael.fellmann@uni-rostock.de
[2] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Stuhlsatzenhausweg 3, 66123 Saarbrücken, <Vorname>.<Nachname>@dfki.de
[3] Karlsruher Institut für Technologie (KIT), Institut AIFB, KIT-Campus Süd, 76128 Karlsruhe, <Vorname>.<Nachname>@kit.edu
[4] Universität Duisburg-Essen, Gerlingstraße 16, 45127 Essen, michael.striewe@paluno.uni-due.de

eröffnet die Möglichkeit der statistischen Auswertung und damit verbunden z. B. die Identifikation häufig auftretender Modellierungsfehler. Dies könnte konstruktiv genutzt werden, um Rückschlüsse auf die eigene Lehre zu ziehen und diese ggf. anzupassen [US16].

Für die Bewertung von im Lehr- und Lernkontext von Studierenden erstellten Modellen gelten nicht unbedingt die in der Literatur zahlreich beschriebenen Qualitätsanforderungen an Modelle (z. B. [LSS94, OBS12]), die für Unternehmen relevant sind. Idealerweise werden auch Lernziele der korrespondierenden Veranstaltung berücksichtigt. Herausforderungen bei der Entwicklung (teil-)automatisierter Anwendungen für die IT-gestützte Bewertung liegen zusätzlich darin, dass einige Aspekte für die automatisierte Analyse von Modellen nur schwer zugänglich sind, wie z. B. die Erkennung von semantischen Fehlern oder die korrekte Interpretation von Elementbeschriftungen. Andererseits werden automatisierten Bewertungsverfahren im Kontext Lehre durch das Vorliegen von Musterlösung(en) und Aufgabentext bzw. Domänenbeschreibung wertvolle Zusatzinformationen zur Verfügung gestellt, die bislang unzureichend genutzt werden. So existieren derzeit nur vereinzelt veröffentlichte Arbeiten auf diesem Gebiet, z. B. erste Ansätze für die automatisierte Bewertung von Modellen in Form von ER- und UML-Diagrammen [STW13, VP14].

Vor diesem Hintergrund wurde im Rahmen des Workshops zur Modellierung in der Hochschullehre (MoHoL) 2016[5] ein Testdatensatz mit Lösungen zu einer gegebenen Modellierungsaufgabe erstellt und die Bewertung dieser Lösungen als Wettbewerb ausgerufen. Zu diesem Wettbewerb wurden zwei Beiträge für die automatisierte Bewertung eingereicht. In diesem Artikel werden zunächst Bewertungskriterien aufgestellt und dann die Ergebnisse der beiden automatisierten Bewertungsansätze aus dem Wettbewerb mit den Ergebnissen einer manuellen Bewertung verglichen. So kann ermittelt werden, welche Bewertungskriterien mit den automatisierten Bewertungsansätzen bereits geprüft werden können und in welchen Bereichen Herausforderungen für die Weiterentwicklung liegen.

Der weitere Artikel ist folgendermaßen aufgebaut: In Kapitel 2 werden zunächst die betrachtete Modellierungsaufgabe, der Testdatensatz sowie Bewertungskriterien und das manuelle Bewertungsschema beschrieben. Die automatisierten Ansätze werden in Kapitel 3 vorgestellt. Ein Vergleich der manuellen Bewertung mit den automatisierten Bewertungsergebnissen stellt daraufhin den Kern des Artikels dar (Kapitel 4). Abschließend werden in Kapitel 5 die Erkenntnisse des Artikels zusammengefasst und ein Ausblick auf zukünftigen Forschungsbedarf gegeben.

2 Evaluationsszenario

Eine wesentliche Kompetenz, welche Studierende in Lehrveranstaltungen zur Modellierung erwerben sollen, ist die Bildung von Modellen (vgl. [Gl08, These 6]). Um diese Kompetenz zu prüfen, wird typischerweise verlangt, zu einer gegebenen textuellen Beschreibung ein grafisches Modell zu erstellen – was in diesem Beitrag als Modellierungsaufgabe bezeichnet wird. Die ausgewählte Modellierungsaufgabe wurde aus einer Klausur einer Vorlesung zur Modellierung von Geschäftsprozessen für Studierende des Wirtschafts-

[5] Workshop zur Modellierung in der Hochschullehre (MoHoL): `http://butler.aifb.kit.edu/MoHoL/`

ingenieurwesens und der Informationswirtschaft am *Karlsruher Institut für Technologie* (KIT) entnommen. Die Aufgabenstellung sieht vor, dass zu einer gegebenen Prozessbeschreibung ein Modell in Form einer Ereignisgesteuerten Prozesskette (EPK) [KNS92] erstellt wird. Der vollständige Aufgabentext ist im Folgenden wiedergegeben.

> *Model the processing of customer inquiries as described in the following using the EPC method: As soon as an inquiry from a customer is received, the feasibility is checked first. This check leads to three different results: either the request is not feasible, possibly feasible or feasible. If the request is not feasible, a rejection of the request is created and the customer is informed subsequently. If the request is possibly feasible, a clarification is brought about. The result of the clarification may be either positive or negative. In case of a negative clarification, a rejection is created and the customer is informed subsequently. In case of a positive clarification, an offer is created and the customer is informed subsequently. If the outcome of the feasibility check showed that the request is feasible, an offer is created and the customer is informed subsequently.*

Ein wichtiges Charakteristikum in der Prozessmodellierung liegt grundsätzlich darin, dass nicht nur ein mögliches Lösungsmodell existiert, sondern unterschiedliche Modelle denselben Sachverhalt korrekt abbilden können. Bei Modellierungsaufgaben, die zu einem künstlichen Sachverhalt gestellt werden, kann man der Varianz in den Lösungen jedoch weitestgehend entgegenwirken, indem die gegebene Prozessbeschreibung so detailliert ist, dass für die Modellierung nur ein geringer oder gar kein Freiheitsgrad herrscht. Für die in diesem Artikel verwendete Modellierungsaufgabe ist es daher möglich, eine Musterlösung zu erstellen, welche den durch den Aufgabentext recht eindeutig definierten Kontrollfluss abbildet (siehe Abb. 1). Darüber hinaus können Divergenzen zur Musterlösung jedoch auch weitere Aspekte des Modells betreffen: Für die Beschriftung und die Anordnung der visuellen Modellelemente kann es unzählige verschiedene – aber korrekte – Varianten geben. Aus diesem Grund stellt die Bewertung unterschiedlichster Lösungen für eine Modellierungsaufgabe eine besondere Herausforderung für automatisierte Ansätze dar.

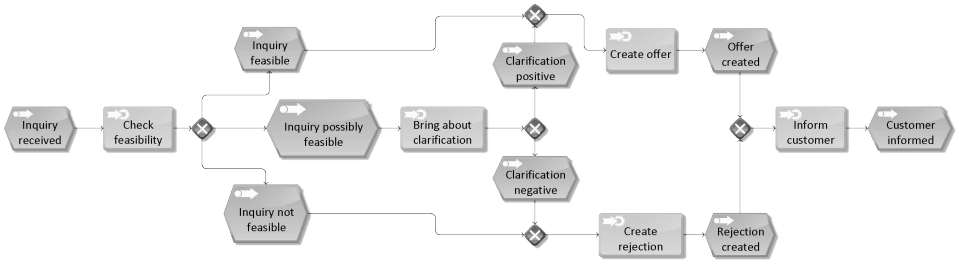

Abb. 1: Musterlösung zur verwendeten Modellierungsaufgabe

2.1 Bewertungskriterien

In der Literatur existiert eine Reihe von Vorschriften und Richtlinien, die bei der Modellierung eingehalten werden sollen. Der Übersichtsartikel von Fellmann et al. [Fe13] liefert eine Zusammenstellung verschiedener Qualitätsaspekte bzw. Bewertungskriterien, die bei der Modellierung von EPK eine Rolle spielen. Die für die hier behandelte Modellierungs-

aufgabe relevanten Kriterien werden im Folgenden nach verschiedenen Qualitätsaspekten gruppiert aufgelistet, wobei zusätzliche relevante Aspekte ergänzt werden.

I Syntaktische Qualität

SYN01 Existenz von jeweils mindestens einem Start- und einem Endereignis [Fe13, Abs. 3.2.2, Regel 1].

SYN02 Auf eine Funktion darf immer nur ein Ereignis folgen und vice versa, mit Ausnahme des Start- und Endereignisses [Fe13, Abs. 3.2.2, Regel 2].

SYN03 Bei der Verwendung logischer Operatoren muss der Join-Operator vom gleichen Typ sein wie der Split-Operator [Fe13, Abs. 3.2.2, Regel 4].

SYN04 Logische Operatoren besitzen stets genau eine Eingangskante und mehrere Ausgangskanten oder vice versa [Fe13, Abs. 3.2.2, Regel 5].

SYN05 Ereignisse und Funktionen besitzen stets genau eine Eingangs- und eine Ausgangskante [Fe13, Abs. 3.2.2, Regel 6].

SYN06 Auf ein Ereignis darf kein XOR- oder OR-Operator folgen, da ein Ereignis keine Entscheidungskompetenz besitzt [Fe13, Abs. 3.2.2, Regel 7].

II Semantische Qualität

SEM01 Eine EPK ist sprachlich korrekt, wenn jede Elementbeschriftung des Modells einer unbestimmten aber konsistent eingesetzten Konvention entspricht (modifiziert im Vergleich zu [Fe13, Abs. 3.3.2], das den Einsatz einer bestimmten Konvention fordert).

SEM02 Ein Modell ist semantisch korrekt, wenn alle Aussagen des Modells korrekt und relevant in Bezug auf die Domäne sind [LSS94].

SEM03 Ein Modell ist semantisch vollständig, wenn durch das Modell alle in der Domäne relevanten Aussagen ausgedrückt werden [LSS94].

SEM04 Der Abstraktionsgrad soll so gewählt werden, dass eine Aktivität bzw. ein Zustand aus der Domäne genau einer Funktion bzw. einem Ereignis entspricht.

III Pragmatische Qualität

PRA01 Das Modell darf keine sich überschneidenden Kanten besitzen [Fe13, Abs. 3.4, Regel 1].

PRA02 Der Kontrollfluss des Modells darf keinen Richtungswechsel aufweisen [Fe13, Abs. 3.4, Regel 2].

2.2 Testdatensatz und manuelles Bewertungsschema

Die Auswahl der Lösungen im Testdatensatz soll eine große Bandbreite an möglichen Ausprägungen abdecken. Er enthält insbesondere Lösungen mit Modellierungsstilen und Fehlern, die häufig in den von den Autoren des vorliegenden Beitrags analysierten Klausurlösungen auftreten. Für den Wettbewerb wurden ursprünglich zehn Lösungen zusammengestellt, davon acht synthetische und zwei reale Lösungen. Für diesen Beitrag wurde der Datensatz um zehn weitere reale Lösungen auf insgesamt 20 Lösungen erweitert, um insbesondere die Anwendungsnähe zu demonstrieren. Bei den acht synthetischen

Kriterium	Kurzbeschreibung	Abzug	additiv
(SYN01)	Existenz von Start-/Endereignis	2	ja
(SYN02)	Alternieren von Funktion/Ereignis	2	ja
(SYN03)	Split- und Join-Operator vom gleichen Typ	2	ja
(SYN04)	Anzahl Ein-/Ausgangskanten von Operatoren	2	ja
(SYN05)	Anzahl Ein-/Ausgangskanten von Funktion/Ereignis	2	ja
(SYN06)	Fehlende Entscheidungskompetenz XOR/OR-Operator	2	ja
(SEM01)	Sprachliche Korrektheit	2	nein
(SEM02)	Semantische Korrektheit	2	ja
(SEM03)	Semantische Vollständigkeit	2	ja
(SEM04)	Abstraktionsgrad und atomare Darstellung	1	ja
(PRA01)	Kein Überschneiden von Kanten	2	nein
(PRA02)	Kein Richtungswechsel des Kontrollflusses	2	nein

Tab. 1: Bewertungsschema für die EPK-Modellierungsaufgabe

Lösungen handelt es sich um Modifikationen der Musterlösung, die in bestimmten Aspekten von den in Abschnitt 2.1 vorgestellten Bewertungskriterien abweichen, um erkennen zu können, ob die automatisierten Ansätze diese Abweichungen erfolgreich identifizieren. Die zwölf realen Lösungen wurden aus vorangegangenen schriftlichen Klausuren ausgewählt und nachträglich digitalisiert, um die automatisierte Analyse zu ermöglichen. Dabei wurde großer Wert darauf gelegt, die Lösungen bei der Digitalisierung nicht zu verändern. D. h. auch das Layout sowie die Modellbeschriftungen wurden – soweit das eingesetzte Modellierungswerkzeug[6] dies zuließ – originalgetreu umgesetzt. Ebenso wurde die Musterlösung dem Testdatensatz hinzugefügt. Alle Lösungen wurden im Anschluss in das XML-basierte Austauschformat EPML[7] überführt. Der vollständige Testdatensatz ist frei verfügbar und kann heruntergeladen werden.[8]

Als Grundlage für den Vergleich der Bewertungen wurde für die ermittelten Kriterien ein Bewertungsschema in Form eines Abzugsverfahrens erstellt, das die Qualität aller Lösungen auf das ganzzahlige Intervall $[0, 20]$ abbilden soll. Dazu wird bei Nichterfüllung von Kriterien eine bestimmte Anzahl von Punkten von der maximal erreichbaren Punktzahl abgezogen. Tabelle 1 gibt an, wie viele Punkte jeweils abgezogen werden, wenn ein Kriterium nicht erfüllt wird. Dabei bestimmt der Eintrag in der Spalte *additiv*, ob bei mehrfacher Nicht-Erfüllung auch ein mehrfacher Punktabzug stattfindet. Es bleibt anzumerken, dass sowohl die Auswahl als auch die Gewichtung der unterschiedlichen Bewertungskriterien subjektive Präferenzen der jeweiligen Lehrperson widerspiegeln. Bei der Erstellung eines Bewertungsschemas muss stets eine Reihe individueller Entscheidungen getroffen werden, z. B. wie stark syntaktische Mängel des Modells gewichtet werden oder wie mit augenscheinlichen Flüchtigkeits- oder Folgefehlern umgegangen wird.

[6] ARIS Community Edition, `http://www.ariscommunity.com/aris-express`
[7] EPC Markup Language (EPML), `http://www.mendling.com/EPML/`
[8] Testdatensatz zum Download: `http://butler.aifb.kit.edu/DeLFI2016_Testdatensatz.zip`

3 Automatisierte Bewertung

Zur automatisierten Bewertung der im Testdatensatz enthaltenen Lösungen beteiligten sich die Werkzeuge *RefMod-Miner* und *JACK* am Wettbewerb. Im Folgenden werden die grundlegenden Ansätze der beiden Werkzeuge sowie die Adressierung der verschiedenen Bewertungskriterien vorgestellt.

3.1 RefMod-Miner

Am *Institut für Wirtschaftsinformatik* (IWi) im *Deutschen Forschungszentrum für Künstliche Intelligenz* (DFKI) und an der *Universität des Saarlandes* wird seit einigen Jahren der Forschungsprototyp *RefMod-Miner* entwickelt, welcher die Analyse von Geschäftsprozessmodellen sowie die induktive Entwicklung von Referenzprozessmodellen ermöglicht. Die implementierten Techniken, u. a. zur automatisierten Erkennung von Korrespondenzen zwischen Prozessmodellen, zur Ähnlichkeitsanalyse, zur Ableitung möglicher Ausführungspfade oder auch zur Berechnung und Analyse von Modellmetriken stellen eine vielversprechende Basis für eine automatisierte Bewertung von Modellierungsaufgaben dar. Ein entsprechendes Konzept wurde im Rahmen des Wettbewerbs entwickelt, prototypisch implementiert [Th16] und öffentlich im *RefMod-Miner as a Service*[9] bereitgestellt.

Zur Beurteilung der syntaktischen Qualität existieren bisher sieben unterschiedliche generische Modellierungsregeln, welche die in Abschnitt 2.1 vorgestellten Regeln, mit Ausnahme des Kriteriums SYN03, vollständig abbilden. Vor dem Hintergrund unterschiedlicher Auffassungen über eine korrekte EPK sind die bei der Bewertung anzuwendenden Regeln frei wählbar. Grundsätzlich wird jede Regelverletzung neben der resultierenden Bewertung in natürlicher Sprache beschrieben.

Während zur Beurteilung der syntaktischen Qualität demnach – bis auf die Auswahl der Bewertungskriterien – kein manueller Aufwand entsteht, basiert die Beurteilung der semantischen Qualität auf einer Musterlösung, welche exakt den in der textuellen Prozessbeschreibung dargelegten Inhalt abbildet. Es ist deshalb notwendig, die Lösungen inhaltlich in einer Weise interpretieren zu können, die einen Abgleich zur Musterlösung zulässt. Im Prototyp wird dazu das Verfahren RefMod-Mine/NHCM [An15] zur automatisierten Identifikation von Korrespondenzen zwischen Prozessmodellen (engl. *Process Model Matching*) verwendet. Davon ausgehend kann das Kriterium SEM03 durch die Precision/Recall-Werte der in der studentischen Lösung enthaltenen Knoten im Verhältnis zu den in der Musterlösung enthaltenen Knoten bestimmt werden. Zur Beurteilung des Kriteriums SEM02 werden ausgehend von den bekannten Korrespondenzen alle möglichen Ausführungspfade berechnet und dadurch ein Abgleich der Zustandsräume von studentischen Lösungen mit der Musterlösung ermöglicht. Die Quantifizierung des Kriteriums SEM02 erfolgt schließlich über den F-Measure, welcher zusätzlich durch die Länge der längsten Teilpfade gewichtet wird. Die Beurteilung des Kriteriums SEM01 ist grundsätzlich durch eine automatische Labeltypisierung im RefMod-Miner möglich,

[9] RefMod-Miner as a Service: `http://rmm.dfki.de`

wurde jedoch im vorliegenden Kontext bisher nicht umgesetzt. Die pragmatische Qualität wie sie im vorliegenden Beitrag definiert ist, bleibt im aktuellen Entwicklungsstand unberücksichtigt, da das Werkzeug Layoutinformationen nicht verarbeitet. Gleichwohl können bereits weitere pragmatische Aspekte beurteilt werden: (1) Verständlichkeit von Prozessmodellen sowie (2) ein Indikator für Täuschungversuche.

Es existieren verschiedene modellbezogene Faktoren, die die Verständlichkeit von Prozessmodellen (1) beeinflussen [HFL14]. In der Literatur konnte eine Vielzahl positiver und negativer Einflüsse auf die Modellverständlichkeit (1) abhängig von bestimmten Metriken wie beispielsweise der Kanten- und Konnektorenanzahl, der Kontrollflusskomplexität oder des Konnektivitätskoeffizienten empirisch nachgewiesen werden [Th16]. Diese Einflüsse werden analysiert und über das Bewertungsergebnis als Feedback zurückgegeben. Ein Indikator für Täuschungsversuche (2) wurde durch die Anwendung der Graph-Editier-Distanz auf die Abbildung identisch beschrifteter Knoten über alle Lösungen hinweg realisiert. Dieser Indikator kann als Ausgangspunkt für eine weitere Prüfung hinsichtlich einer möglichen Täuschung herangezogen werden.

Die Gewichtung der einzelnen Bewertungskriterien kann frei konfiguriert werden, ebenso wie die Maximalpunktzahl. Auf diese Weise konnte das in Abschnitt 2 vorgestellte Bewertungsschema mit Ausnahme der nicht unterstützten Aspekte in nur wenigen Minuten umgesetzt werden. Hinzu kommt dabei die Erstellung der Musterlösung.

3.2 JACK

Die an der *Universität Duisburg-Essen* entwickelte automatische Bewertung der Aufgaben basiert auf dem eAssessment-System *JACK*[10], für das der Prototyp einer Komponente zur Analyse von EPKs entwickelt wurde. Diese Komponente ist in der Lage, EPKs im EPML-Dateiformat einzulesen. Lehrende können für eine Aufgabe Regeln definieren, die erwünschte oder unerwünschte Modellstrukturen oder Inhalte in Lösungen beschreiben. Jede Regel enthält ein textuelles Feedback, das bei Verletzung der Regel ausgegeben wird. Ferner ist es möglich, Regeln unterschiedlich stark zu gewichten sowie einzustellen, ob mehrfache Verletzungen derselben Regel auch mehrfach gewertet werden sollen.

Das Bewertungsschema aus Abschnitt 2 kann auf dieser technischen Grundlage zu großen Teilen, aber nicht vollständig umzusetzen. Eine Einschränkung betrifft die pragmatische Qualität (PRA01 und PRA02), die wie oben bereits festgestellt nicht geprüft werden kann, da Dateien im EPML-Format keine Layoutinformationen enthalten. Eine weitere Einschränkung betrifft Kriterium SEM01, das sich nur heuristisch überprüfen lässt. Hierfür wird ein sehr einfacher statistischer Ansatz verwendet, der das Verhältnis zwischen der Anzahl unterschiedlicher Worte und der Anzahl der Knoten im Diagramm berücksichtigt.

Alle anderen Aspekte des Bewertungsschemas lassen sich durch entsprechende Regeln exakt abbilden. Die Regeln der syntaktischen Kriterien sowie einige der semantischen Kriterien sind universell verwendbar, da sie unabhängig von der Aufgabenstellung for-

[10] JACK: `http://www.s3.uni-duisburg-essen.de/jack/`

muliert werden können. Ein Beispiel für eine Regel für ein syntaktisches Kriterium ist in Listing 1 angegeben. Die übrigen Regeln sind aufgabenspezifisch, da sie konkret auf Begriffe aus der Aufgabenstellung Bezug nehmen. Bei der semantischen Vollständigkeit der Lösung SEM03 muss für jeden erwarteten Aspekt aus der Aufgabenstellung eine eigene Regel formuliert werden, wobei innerhalb einer solchen Regel auch Formulierungsvarianten berücksichtigt werden können. Eine automatische Berücksichtigung von Tippfehlern ist im Prototypen jedoch nicht möglich. Diese Schwäche betrifft auch Kriterium SEM02, bei dessen Umsetzung unter anderem alle vorhandenen Elemente gegen erwartete Texte aus der Aufgabenstellung geprüft werden, sowie Kriterium SEM04, bei dessen Umsetzung unter anderem gleichnamige Elemente gesucht werden.

Der komplette Regelsatz zur Abbildung des Bewertungsschemas umfasst 21 Regeln, die jeweils Gewicht 1 oder 2 haben, so dass das Gesamtgewicht 40 Punkte beträgt. Die erreichte Punktzahl wird daher am Ende auf die gewünschte Skala von 0 bis 20 umgerechnet. Erstellung, Verifikation und Nachbesserung des Regelsatzes erforderten in Summe etwas mehr als einen Arbeitstag, wovon etwa die Hälfte auf aufgabenspezifische Regeln entfiel.

```
<rule type="absence" points="2" multi="true">
   <query> from x: V{Event}, y: V{Connector}
      with (x --> y) and (type(y)="Or" or type(y)="Xor")
      report x.name as "xName", type(y) as "yName" end</query>
   <feedback prefix="Fehlerhafte Grundstruktur">Dem Ereignis "{xName}"
       folgt ein {yName}-Operator. Dies ist nicht erwünscht, da
       Ereignisse keine Entscheidungen treffen können.</feedback>
</rule>
```

List. 1: Prüfregel für Kriterium SYN06. Die Regel ist vom Typ *absence*, d. h. es wird in der Anfrage eine unerwünschte Struktur beschrieben. Die Gewichtung der Regel ist 2 und sie ist mehrfach anwendbar. Die Anfrage verwendet die Graphabfragesprache GReQL, um eine Struktur zu beschreiben, in der der Operator *y* unmittelbar auf das Ereignis *x* folgt und vom Typ OR oder XOR ist.

4 Vergleich der Bewertungsergebnisse

Die Ergebnisse der manuellen Bewertung sowie der automatisierten Ansätze *JACK* und *RefMod-Miner* sind in Tabelle 2 zusammengefasst. Dabei wurde jeweils auch die Abweichung der automatisierten Ansätze zur manuellen Bewertung berechnet und aufsummiert.

Zunächst ist festzuhalten, dass die beiden automatisierten Ansätze im Schnitt 2,15 Punkte (*RefMod-Miner*) bzw. 1,90 Punkte (*JACK*) und damit deutlich von der manuellen Bewertung abweichen. Gleichwohl werden 7 (*RefMod-Miner*) bzw. 9 (*JACK*) exakte Übereinstimmungen mit der manuellen Bewertung erreicht. Darunter befindet sich auch die Musterlösung, die stets die volle Punktzahl erhält. Ferner erreicht *RefMod-Miner* 6 Bewertungen, die um maximal zwei Punkte von der manuellen Bewertung abweichen, während dies bei *JACK* in 5 Fällen zutrifft. In 8 (*RefMod-Miner*) bzw. 7 (*JACK*) Fällen liegt die Abweichung über 2 Punkten. Die Unterschiede zwischen den automatisierten Ansätze sind etwas geringer: Es werden 9 Lösungen identisch bewertet, weitere 9 mit einer Abweichung von maximal 2 Punkten und nur in 3 Fällen liegt die Abweichung höher. Die durchschnittliche Abweichung der beiden Ansätze zueinander liegt insgesamt bei 1,75.

		Manuell		RefMod-Miner			JACK			
Lsg.	Typ	Bew.	Verletzte Kriterien	Bew.	Verletzte Kriterien	Abw.	Bew.	Verletzte Kriterien	Abw.	Abw. aut.
00	M	20	-	20	-	(0)	20	-	(0)	(0)
01	S	14	SEM04 (6)	14	SEM04 (6)	(0)	14	SEM04 (6)	(0)	(0)
02	S	18	SEM04 (2)	18	SEM04 (2)	(0)	18	SEM04 (2)	(0)	(0)
03	S	16	SYN05 (2)	16	SYN05 (2)	(0)	16	SYN05 (2)	(0)	(0)
04	S	16	SYN03 (2)	16	SEM02 (2)	(0)	16	SYN03 (2)	(0)	(0)
05	R	08	SYN02, SYN03, SYN05, SEM02, SEM03, PRA02	10	SYN05 (2), SEM02, SEM03 (2)	(2)	12	SYN05, SEM02 (2), SEM03	(4)	(2)
06	R	13	SYN03 (2), SEM04 (3)	09	SEM02 (3), SEM03, SEM04 (3)	(4)	11	SYN03 (2), SEM01, SEM04 (3)	(2)	(2)
07	S	12	SEM02 (4)	10	SEM02 (4), SEM03 (2)	(2)	12	SYN03 (2), SEM02 (2)	(0)	(2)
08	S	14	SEM02 (2), SEM03	14	SEM02, SEM03 (2)	(0)	16	SYN03, SEM02	(2)	(2)
09	S	14	SEM01, PRA01, PRA02	20	-	(6)	18	SEM02	(4)	(2)
10	S	12	SYN01 (2),SYN02 (2)	12	SYN02 (2), SYN05 (2)	(0)	12	SYN01 (2), SYN02 (2)	(0)	(0)
11	R	14	SEM04 (6)	15	SEM04 (5)	(1)	14	SEM04 (6)	(0)	(1)
12	R	08	SYN05 (3), SEM01, PRA01, PRA02	14	SYN05 (3)	(6)	14	SYN05 (3)	(6)	(0)
13	R	14	SYN05, SEM01, SEM04 (2)	18	SYN05	(4)	18	SYN05	(4)	(0)
14	R	12	SEM01, SEM04 (6)	13	SEM04 (7)	(1)	11	SEM02, SEM04 (7)	(1)	(2)
15	R	16	SEM01, PRA02	14	SEM02, SEM03 (2)	(2)	18	SEM03	(2)	(4)
16	R	05	SYN02, SEM01, SEM02 (3), SEM04, PRA01, PRA02	08	SYN02, SEM02 (3), SEM03, SEM04 (2)	(3)	06	SYN02, SYN03 (4), SEM02, SEM04 (2)	(1)	(2)
17	R	09	SYN01, SYN05 (2), SEM01, SEM02, SEM04	06	SYN05 (4), SEM02, SEM03 (2)	(3)	14	SYN01, SYN05 (2)	(5)	(8)
18	R	12	SYN05 (2), SEM01, PRA02	14	SYN05 (2), SEM03	(2)	12	SYN05 (2), SEM02, SEM03	(0)	(2)
19	R	16	SEM01, SEM04 (2)	20	-	(4)	20	-	(4)	(0)
20	R	09	SYN05, SEM02 (3), SEM04	06	SYN05 (2), SEM02 (3), SEM03 (2)	(3)	12	SYN05 (2), SEM02 (2)	(3)	(6)
					Summe Abw.:	(43)			(38)	
					Durchschnittliche Abw.:	(2,15)			(1,90)	

Tab. 2: Ergebnisse der Anwendung des manuellen Bewertungsschemas sowie der automatisierten Ansätze auf den Testdatensatz (siehe jeweils Spalte *Bew.*). Abkürzungen in Spalte *Typ*: R: Real, S: Synthetisch, M: Musterlösung/Referenz. Die Zahl in Klammern hinter einem Kriterium in der Spalte *Verletzte Kriterien* gibt die Anzahl des Auftretens an. Die Ermittlung der durchschnittlichen Abweichung berücksichtigt 20 Lösungen ohne die Referenzlösung 00. In der letzten Spalte ist die absolute Abweichung der beiden automatisierten Ansätze angegeben.

Eine genauere Differenzierung der Unterschiede zur manuellen Bewertung nach Lösungstypen ergibt, dass beide automatischen Ansätze auf den acht synthetischen Lösungen mit einer mittleren Abweichung von 1,00 (*RefMod-Miner*) bzw. 0,75 (*JACK*) deutlich besser abschneiden als auf den realen Lösungen mit einer mittleren Abweichung von 2,92 (*RefMod-Miner*) bzw. 2,67 (*JACK*). Die Zuverlässigkeit scheint damit auf den ersten Blick bei den synthetischen Lösungen höher zu sein, was allerdings dadurch relativiert werden muss, dass die in den realen Lösungen enthaltenen Fehler naturgemäß deutlich heterogener und ausgeprägter sind. Auch ist zu beachten, dass diese Werte aufgrund der geringen Größe der Stichprobe nicht statistisch signifikant sind.

Mit Ausnahme von fünf Fällen bei *RefMod-Miner* und zwei Fällen bei *JACK* fällt die Bewertung der automatischen Ansätze durchweg positiver aus als die manuelle Bewertung. Erklärbar ist dies in den Bewertungskriterien, welche entsprechend Abschnitt 3 tatsächlich durch *JACK* und *RefMod-Miner* überprüft werden. So entfallen bei beiden Werkzeugen die pragmatischen Aspekte vollständig und die Kriterien SYN03 und SEM01 werden nur durch JACK berücksichtigt. Vernachlässigt man die pragmatische Qualität auch in der manuellen Bewertung, reduzieren sich die Abweichungen zur manuellen Bewertung auf 1,55 (*RefMod-Miner*) bzw. 1,50 (*JACK*) und liegen damit auf einem ähnlichen Niveau wie die Abweichungen der Werkzeuge untereinander. Die Unterschiede betreffen dann zudem fast ausschließlich reale Lösungen, wobei sich beide Werkzeuge auch dort leicht auf eine durchschnittliche Abweichung von 2,25 (*RefMod-Miner*) bzw. 2,33 (*JACK*) verbessern. Dass die Verbesserung nicht stärker ausfällt liegt daran, dass in einigen Fällen andere Kriterien stärker gewertet werden als in der manuellen Bewertung. So wertet *JACK* bei Lösung 16 eine vierfache Verletzung von SYN03, die durch zwei fehlerhafte Join-Konnektoren ausgelöst wird und in der manuellen Bewertung nur als doppelte Verletzung von SEM02 gewertet wird.

Die fünf Fälle, in denen *RefMod-Miner* negativer bewertet als die manuelle Bewertung, lassen sich wie folgt erklären: (1) In der manuellen Bewertung wurden Fehler teilweise einfach bewertet, obwohl gleichzeitig gegen zwei Bewertungsaspekte verstoßen wurde (z. B. Lösung 06), diese wurden im *RefMod-Miner* doppelt gewertet; (2) Die automatisierte Identifizierung korrespondierender Knoten führt in wenigen Fällen zu Fehlinterpretation, die in Folge zu erhöhten Abzügen in SEM02 und SEM03 führen (Lösungen 07, 15, 17 und 20). Die beiden negativeren Bewertungen durch *JACK* haben zwei individuelle Ursachen: Bei Lösung 06 wertet das heuristische Verfahren für SEM01 strenger als die manuelle Bewertung. Bei Lösung 14 wird eine zusätzliche Verletzung von SEM04 gezählt, da *JACK* bei diesem Kriterium nur wenig sprachliche Toleranzen zulässt.

Es gibt jedoch auch Fälle, in denen die automatischen Ansätze deutlich positiver werten als die manuelle Bewertung, ohne dass dies an nicht geprüften Kriterien liegt. Exemplarisch sind hier die Lösungen 13, 17, 19 und 20 zu nennen, in denen die manuelle Bewertung jeweils eine Verletzung von SEM04 wertet, die durch die automatisierten Ansätze nicht erkannt wird. Dies liegt daran, dass zur Bewertung dieses Kriteriums Elemente der Modelle sowohl untereinander als auch mit der Musterlösung verglichen werden müssen. Beide automatisierten Ansätze vergleichen jedoch Elemente entweder mit der Musterlösung oder untereinander und sind daher aktuell nicht darauf ausgerichtet, Dopplungen zu finden.

Schließlich ist noch zu beobachten, dass es Fälle gibt, in denen eine exakte oder weitgehende Übereinstimmung der Punktzahl erreicht wird, jedoch andere Kriterien als verletzt gewertet werden. Dies liegt an einem gewissen Interpretationsspielraum, der durch die Kriterien sowohl für die manuelle als auch die automatische Bewertung gegeben ist. Bei Lösung 04 wird beispielsweise manuell ein unpassender Join-Konnektor als syntaktischer Fehler gewertet, während *RefMod-Miner* diesen als inhaltlich falsch betrachtet und daher einen semantischen Fehler wertet. Bei Lösung 14 wird manuell eine sprachliche Inkonsistenz gewertet, während *JACK* eine semantische Inkorrektheit sieht. Diese Abweichungen sind insbesondere mit Blick auf die automatische Feedback-Erzeugung relevant, wenn nicht nur eine Note, sondern auch eine textuelle Erläuterung ausgegeben werden soll.

5 Fazit

Die untersuchten automatisierten Bewertungsansätze für Modellierungsaufgaben weisen insgesamt vielversprechende Potenziale für einen Einsatz in der (Hochschul-)Lehre auf. Die syntaktischen Kritieren werden bereits vollautomatisiert zuverlässig geprüft, während je nach Ansatz und Bewertungskriterium weitere Herausforderungen existieren.

Eine Herausforderung beim Ansatz des *RefMod-Miners* besteht insbesondere in der automatischen Identifizierung von Korrespondenzen zwischen studentischen Lösungen und der Musterlösung. Zwar wurden in der nahen Vergangenheit vielfältige automatisierte Verfahren für diese Aufgabe entwickelt, allerdings sind diese bei weitem noch nicht zuverlässig genug. In diesem Zusammenhang bietet sich deshalb ein hybrides bzw. teilautomatisiertes Vorgehen an, in dem Korrespondenzen manuell nachbearbeitet werden, um die weitere Bewertung auf Basis einer gesicherten Datenbasis automatisiert vorzunehmen. Ein Vorteil der angewandten Methode ist die Unabhängigkeit der Bewertung von der konkreten Aufgabenstellung. Vergleichbare Techniken wurden in *JACK* bisher nur für andere Diagrammtypen angewendet [SG14] und müssten im Rahmen zukünftiger Arbeiten übertragen werden, da der regelbasierte Ansatz einen hohen Aufwand beim Erstellen aufgabenspezifischer Regeln erfordert. In beiden Ansätzen stellt insbesondere die Identifikation korrespondierender Labels eine wichtige Herausforderung dar, die weiterer Forschung bedarf. Die Berücksichtigung der pragmatischen Aspekte erfordert ebenfalls weitere Arbeiten und sowohl eine Erweiterung der Werkzeuge als auch die Wahl eines anderen Dateiformats, damit die für die Bewertung der pragmatischen Aspekte notwendigen Informationen überhaupt vorliegen.

Auf der Seite der Evaluation ist die Weiterführung des Vergleichs mit größeren Datensätzen wünschenswert, um statistisch belastbarere Aussagen zu erhalten. Dies schließt gezielte Evaluierungen mit Blick auf einzelne Kriterien oder Bewertungsverfahren ebenso ein wie die Analyse der Inter-Rater-Reliability in Relation zu mehreren manuellen Bewertern auf der Suche nach einem Goldstandard für die Bewertung. Gegebebenfalls ist auch die Einführung weiterer Bewertungskriterien (z. B. basierend auf Metriken) notwendig, um die Ansprüche manueller Bewertungen komplett abzudecken. Nicht zuletzt ist es auch möglich, die bisherigen Ansätze auf weitere Modellierungssprachen auszuweiten, bei denen vergleichbare Bewertungskritierien Anwendung finden.

Literaturverzeichnis

[An15] Antunes, Goncalo; Bakhshandeh, Marzieh; Borbinha, Jose; Cardoso, Joao; Dadashnia, Sharam; Francescomarino, Chiara Di; Dragoni, Mauro; Fettke, Peter; Gal, Avigdor; Ghidini, Chiara; Hake, Philip; Khiat, Abderrahmane; Klinkmüller, Christopher; Kuss, Elena; Leopold, Henrik; Loos, Peter; Meilicke, Christian; Niesen, Tim; Pesquita, Catia; Péus, Timo; Schoknecht, Andreas; Sheetrit, Eitam; Sonntag, Andreas; Stuckenschmidt, Heiner; Thaler, Tom; Weber, Ingo; Weidlich, Matthias: The Process Model Matching Contest 2015. In: 6th International Workshop on Enterprise Modeling and Information Systems Architectures. S. 127–155, 2015.

[Fe13] Fellmann, Michael; Bittmann, Sebastian; Karhof, Arne; Stolze, Carl; Thomas, Oliver: Do we need a Standard of EPC Modelling? The State of Syntactic, Semantic and Pragmatic Quality. In: 5th International Workshop on Enterprise Modelling and Information Systems Architectures. S. 103–116, 2013.

[Gl08] Glinz, Martin: Modellierung in der Lehre an Hochschulen: Thesen und Erfahrungen. Informatik Spektrum, (31/5):425–434, 2008.

[Ha16] Hafer, Jörg; Matthé, Frederic; Schumann, Wilfried; Wollersheim, Heinz-Werner; Jeremias, Christoph; Grigat, Felix; Schulz, Alexander: Schwerpunktthema: E-Klausuren. Forschung und Lehre, (3):194–209, 2016.

[HFL14] Houy, Constantin; Fettke, Peter; Loos, Peter: On the Theoretical Foundations of Research into the Understandability of Business Process Models. In: 22nd European Conference on Information Systems. S. 1–26, 2014.

[KNS92] Keller, Gerhard; Nüttgens, Markus; Scheer, August-Wilhelm: Semantische Prozessmodellierung auf der Grundlage Ereignisgesteuerter Prozessketten (EPK). Bericht 89, Institut für Wirtschaftsinformatik, 1992.

[LSS94] Lindland, Odd Ivar; Sindre, Guttorm; Sølvberg, Arne: Understanding Quality in Conceptual Modeling. IEEE Software, (11/2):42–49, 1994.

[OBS12] Overhage, Sven; Birkmeier, Dominik; Schlauderer, Sebastian: Qualitätsmerkmale, -metriken und -messverfahren für Geschäftsprozessmodelle - Das 3QM-Framework. Wirtschaftsinformatik, (54/5):217–235, 2012.

[SG14] Striewe, Michael; Goedicke, Michael: Automated Assessment of UML Activity Diagrams. In: Conference on Innovation & Technology in Computer Science Education. S. 336, 2014.

[STW13] Smith, Neil; Thomas, Pete; Waugh, Kevin: Automatic grading of free-form diagrams with label hypernymy. In: Learning and Teaching in Computing and Engineering. IEEE, S. 136–142, 2013.

[Th16] Thaler, Tom; Houy, Constantin; Fettke, Peter; Loos, Peter: Automated Assessment of Process Modeling Exams: Basic Ideas and Prototypical Implementation. In: Workshop zur Modellierung in der Hochschullehre. S. 63–70, 2016.

[US16] Ullrich, Meike; Schoknecht, Andreas: (Business Process) Models from an Educational Perspective. In: 8th Central European Workshop on Services and their Composition Workshop. S. 53–55, 2016.

[VP14] Vachharajani, Vinay; Pareek, Jyoti: A proposed architecture for automated assessment of use case diagrams. International Journal of Computer Applications, (108/4), 2014.