EVALUATION VON NUTZERSIMULATIONEN ANHAND EINES ÄHNLICHKEITSMASSES FÜR DIALOGKORPORA

Stefan Hillmann

stefan.hillmann@tu-berlin.de
Telekom Innovation Laboratories, Technische Universität Berlin

Kurzfassung: Mittels einer Leave-One-Out-Kreuzvalidierung werden Klassifikatoren auf ihre Leistungsfähigkeit bei der Klassifizierung von Dialogen getestet. Die Klassifikation erfolgt mittels Distanzmaßen, welche auf Dialogakt-basierte N-Gramm-Modelle von Dialogen und Dialogkorpora angewendet werden. Die Kreuzvalidierung dient dazu den optimalen Klassifikator für vorhandene Daten zu bestimmen. Die Konfiguration des Klassifikators wird dann dafür genutzt, die Unterschiede zwischen Dialogkorpora zu bestimmen. Die Methode kann u. a. dazu verwendet werden, die Simulation von Dialogverläufen zu bewerten.

1 Einleitung

Mit Hilfe der Simulation von Dialogverläufen können in der Mensch-Maschine-Interaktion Aspekte der Usability von Sprachdialogsystemen automatisiert evaluiert werden. Dazu werden die simulierten Dialoge hinsichtlich relevanter Interaktionsparameter, wie z. B. Aufgabenerfolg, Dialoglänge, Konzeptfehlerrate oder Grad der Abweichung vom optimalen Lösungsweg, untersucht. Anhand der Analyseergebnisse kann die Usability des Systems bewertet werden, ohne dass eine Untersuchung mit realen Nutzern durchgeführt wird.

Voraussetzung dafür ist, dass die simulierten Dialogverläufe den Verläufen aus realen Interaktionen – zumindest hinsichtlich der zu analysierenden Interaktionsparameter – hinreichend ähnlich sind. In dieser Arbeit soll ein Verfahren vorgestellt werden mit dem der Unterschied zwischen zwei Korpora von Dialogen bestimmt werden kann. Für zwei Dialogkorpora, die mit verschiedenen Simulationen erzeugt wurden ist es dann möglich zu entscheiden, welcher einem empirisch erhobenen Dialogkorpus ähnlicher ist. Dies erlaubt eine Bewertung und Ordnung der Simulationswerkzeuge anhand der gemessenen Ähnlichkeit zu realen Dialogen.

In Abschnitt 2 werden zunächst die Grundlagen zur Bildung von N-Gramm-Modellen auf Basis von Dialogakten vorgestellt. Dann werden verschiedene Distanzmaße eingeführt die genutzt werden können, um den Unterschied zwischen zwei N-Gramm-Modellen zu bestimmen. Wie die Distanzmaße objektiv bewertet werden können, indem sie als Klassifikator verwendet werden, und wie sie zum Messen von Unterschieden eingesetzt werden können ist in Abschnitt 3 beschrieben. Die erzielten Resultate werden in Abschnitt 4 vorgestellt. Schließlich werden in Abschnitt 5 die Eigenschaften der Distanzmaße und die Resultate diskutiert und eine Ausblick auf die kommenden Arbeiten gegeben.

2 Methodische Grundlagen

Eine Übersicht verschiedener Ansätzen zur Evaluierung von Nutzersimulationen gibt [7]. Pietquin und Hastie nennen dort u. a. die Nutzung von N-Gramm-Modellen und deren Vergleich mittels Kullback-Leibler Divergenz. In [3] werden auf Basis von Dialogakten N-Gramme gebildet, die jedoch nicht der Evaluation dienen. Vielmehr wird das resultierende N-Gramm-Modell

direkt als Nutzermodell zur Simulation von Dialogverläufen verwendet. Ein weiterer Ansatz zum Vergleich von Nutzersimulationen ist in [10] beschrieben. Williams vergleicht dort allerdings nur die Auftretenswahrscheinlichkeit von Dialogakten, aber betrachtet nicht deren zeitliche Abfolge.

2.1 Erzeugung von N-Gramm-Modellen auf Basis von Dialogakten

Dialoge können als Sequenzen von Dialogakten betrachtet und in Teilsequenzen der Länge n zerlegt werden. Für ein entsprechend annotiertes Dialogkorpus kann auf Basis dieser Teilsequenzen ein N-Gramm-Modell berechnet werden. Ein Dialogschritt teilt sich in die Äußerung eines Dialogpartners und die Antwort des anderen Partners darauf. Bei den hier verwendeten N-Grammen, besteht ein Unigramm (n=1) aus genau einem solchen Schritt; ein Trigramm (n=3) aus 3 Dialogschritten. Ein Trigramm enthält Informationen aus 6 alternierenden Äußerungen (3 pro Dialogteilnehmer), die in ihrer zeitlichen Abfolge angeordnet sind. Für den hier beschriebenen Ansatz ist hervorzuheben, dass die Elemente eines N-Gramms Eigenschaften von einem Dialogakt (Akttyp, verwendete Konzepte) widerspiegeln und nicht direkt aus den konkreten textuellen Äußerungen erzeugt werden.

Das N-Gramm-Modell eines Dialogs oder Dialogkorpus beschreibt für jedes seiner N-Gramme wie häufig es auftritt, wobei im Folgenden immer von der relativen Häufigkeit ausgegangen wird. In diesem Fall, kann ein N-Gramm-Modell als Vektor von rationalen Zahlen dargestellt werden. Die Position eines Wertes im Vektor entspricht einem konkreten N-Gramm, der Wert selber der relativen Häufigkeit dieses N-Gramms im Korpus.

2.2 Glättung eins N-Gramm-Modells

Vor dem Vergleich von zwei N-Gramm-Modellen müssen aus Beiden zwei neue Modelle erstellt werden, welche jeweils die Vereinigungsmenge der N-Gramme aus den ursprünglichen Modellen enthalten. Durch diesen Vorgang kann die Auftretenswahrscheinlichkeit (p) eines N-Gramms in einem der neuen Modelle 0 betragen. Diese 0-Werte müssen durch eine geeignetes Verfahren auf einen Wert $p > 0$ angepasst werden, mit der Bedingung, dass die Summe aller Wahrscheinlichkeiten in einem Modell weiterhin 1 beträgt. Im Rahmen der hier beschriebenen Arbeiten wurde dafür die Laplace Glättung [1] (s. Gl. 1) verwendet.

$$p_\lambda(x_i) = \frac{|x_i| + \lambda}{|X| + \lambda N} \tag{1}$$

$p_\lambda(x_i)$ entspricht der Wahrscheinlichkeit, dass das i-te N-Gramm x_i in dem Modell m auftritt. Dabei ist $|x_i|$ die absolute Häufigkeit von x_i in m, $|X|$ die Anzahl aller N-Gramme in m und N die Anzahl der voneinander verschiedene N-Gramme in m. λ ist ein frei gewählter Parameter größer 0 (z. B. $0,5$).

2.3 Messung der Ähnlichkeit von N-Gramm-Modellen

Mit jedem der nachfolgend vorgestellten Maße kann die Ähnlichkeit von zwei Vektoren P und Q berechnet werden, welche z. B. aus den N-Gramm-Modelle m_P und m_Q bestimmt wurden. Die Dimensionen von P und Q müssen in jedem Fall gleich sein, und Werte an der gleichen Position, sich auf das selbe N-Gramm beziehen.

Die Kosinusähnlichkeit [8] (s. Gl. 2) ist ein geometrisch motiviertes Maß und entspricht dem Winkel zwischen zwei Vektoren. Für sie gilt im Allgemeinen $-1 \leq cd(P||Q) \leq 1$, bei der Anwendung auf N-Gramm-Modelle kann das Intervall jedoch auf $0 \leq cd(P||Q) \leq 1$ eingegrenzt

[1] Auch als Laplace smoothing oder additive smoothing bekannt.

werden. Dies ist möglich, da ein N-Gramm nicht weniger häufig als 0 mal in einem Korpus auftreten kann, weshalb P und Q hier keine negativen Werte enthalten können. Im Folgenden wird die, auf der Kosinusähnlichkeit basierende, Kosinusdistanz (s. Gl. 3) verwendet, bei welcher 0 völlige Übereinstimmung und 1 maximaler Unterschied bedeutet. Die Kosinusdistanz ist kommutativ, d. h. es gilt $cd(P||Q) = cd(Q||P)$.

Die Kullbak-Leibler Divergenz [5] (KLD, s. Gl. 4) berechnet den Unterschied zwischen zwei Wahrscheinlichkeitsverteilungen und wurde bereits, z. B. in [10, S. 39] sowie [7, S. 68], zur Bewertung von Nutzermodellen eingesetzt. Ein entscheidendes Merkmal der KLD ist, dass sie *nicht* kommutativ ist, d. h. $kld(P||Q) = kld(Q||P)$ gilt nicht zwingend.

Um ein echtes Distanzmaß zu erhalten, können die Mittlere Kullback-Leibler Divergenz [7, S. 62] (mKLD, s. Gl. 5) oder auch die Symmetric Kullback-Leibler Divergenz [1, S. 309] (sKLD, s. Gl. 6) verwendet werden, die beide kommutativ sind. Da allerdings gilt dass $sKLD(P||Q) = 2 * mKLD(P||Q)$, wird hier nur die mKLD weiter betrachtet.

Für die KLD gilt $0 \leq kld(P||Q) < \infty$ und je größer $kld(P||Q$ ist, um so unähnlich sind sich P und Q. Beides gilt analog für die mKLD und die sKLD

Die Jensen Differenzen Distanz [9] (JD, s. Gl. 7) bestimmt ebenfalls den Unterschied zwischen zwei Wahrscheinlichkeitsverteilungen. Sie ist ist kommutativ und es gilt $0 \leq jd(P||Q) \leq 1$.

Kosinusähnlichkeit	$cs(P\|\|Q) = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2 \sum_{i=1}^{n} q_i^2}}$	(2)
Kosinusdistanz	$cd(P\|\|Q) = 1 - cs(P\|\|Q)$	(3)
Kullback-Leibler Divergenz (KLD)	$kld(P\|\|Q) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right)$	(4)
Mittlere KLD	$mkl(P\|\|Q) = \frac{kl(P\|\|Q) + kl(Q\|\|P)}{2}$	(5)
Symmetric KLD	$skl(P\|\|Q) = \sum_{i=1}^{n} (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right)$	(6)
Jensen Differenzen Distanz	$jd(P\|\|Q) = \sum_{i=1}^{n} \frac{p_i \ln(p_i) + q_i \ln(q_i)}{2} - \frac{p_i + q_i}{2} \ln\left(\frac{p_i + q_i}{2}\right)$	(7)

3 Bewertung der Distanzmaße und ihre Anwendung

Für die in diesem und den folgenden Abschnitten beschriebenen Arbeiten, wurden Dialoge verwendet die zum einen empirisch erhoben und zum anderen mittels Simulationen erzeugt wurden. Die Simulationen erfolgten mit zwei verschiedenen Aufgabenmodellen welche unterschiedlich realistische Dialoge (im Sinne der Ähnlichkeit zu den Dialogen aus einem empirischen Experiment) erzeugten. Sowohl der empirische Datensatz als auch die beiden Simulation mit unterschiedlichen Aufgabenmodellen sind in [4] beschrieben. Der einzige Unterschied ist, dass in der aktuellen Arbeit auch Dialoge aus der fünften Aufgabenstellung (vgl. [4, S. 25]) der ursprünglichen Studie verwendet wurden.

3.1 Objektive Bewertung der Distanzmaße

In diesem Abschnitt wird beschrieben, wie die oben vorgestellten Distanzmaße durch ihre Leistung als Klassifikator bewertet werden können. Ein Klassifikator wurde jeweils mit einem

Quelle	Kriterium	Kürzel	Anzahl	
			Dialoge	Dialogschritte
Simulation	besseres Aufgabenmodell	sb	1.580	12.456
	schlechteres Aufgabenmodell	sw	1.580	10.297
Experiment	kurze Dialoge	si	49	137
	lange Dialoge	li	49	1.132
	Aufgabenziel erreicht	ts	123	1.286
	Aufgabenziel nicht erreicht	tf	73	715
	Worterkennungsrate 100 %	w1	99	787
	Worterkennungsrate 60 %	w6	17	290

Tabelle 1 - Kriterien anhand derer die Testkorpora zusammengestellt wurden. Aus den Simulationen wurden jeweils alle erzeugten Dialoge verwendet.

konkreten Distanzmaß, einem Wert für n (1–8) und einem Wert für den Glättungsfaktor λ (0,05, 0,25 oder 0,5; s. Abschnitt 2.2) konfiguriert. Jeder dieser Konfigurationen wurde auf einer Reihe von Dialogkorpora getestet, die aus dem emprischen (196 Dialoge mit insgesamt 2001 Dialogschritten) und den simulierten Dialogen extrahiert wurden.

Die Kriterien zur Auswahl von Dialogen sind in Tabelle 1 gelistet und werden nachfolgend kurz beschrieben. In der Tabelle angegebene Kürzel werden auch in den nachfolgenden Tabellen wieder verwendet.

Die Simulation mit dem besseren Aufgabenmodell erzeugte realistischere Dialogverläufe, als die mit dem schlechteren Aufgabenmodell. Zur Ermittlung der kurzen und langen Dialoge wurden alle Dialoge des empirischen Korpus nach ihrer Dialoglänge (Anzahl Dialogschritte) aufsteigend sortiert und jeweils die Ersten und Letzten 25 % der so geordneten Dialoge verwendet. Bei der Auswahl anhand der Erreichung Aufgabenziels (s. [4, S.27]) durch den Nutzer, wurden all Dialoge verwendet und jeweils genau einer der beiden Möglichkeiten (Aufgabenziel erreicht oder nicht erreicht) zugeordnet. Das empirische Korpus enthält Dialoge die bei unterschiedlichen Worterkennungsraten (60 %, 70 %, 80 %, 90 % und 100 %) des Testsystems (BoRIS, [6, S. 241–44 und S. 252–255]) erhoben wurden. In der aktuellen Arbeit wurden alle Dialoge verwendet, die in der damaligen Untersuchung bei einer Worterkennungsrate von 60 % bzw. 100 % erhoben wurden. Das Ziel bei der Auswahl über die Kriterien war es jeweils zwei Teilkorpora von Dialogen zu erhalten, die sich anhand der Dialogverläufe der enthaltenen Dialoge möglichst weit unterscheiden.

Für jede der 96 Kombination aus Distanzmaß, n und λ wurde ein Klassifikator trainiert und mittels Leave-One-Out-Kreuzvalidierung bewertet. Das Training fand jeweils mit den Korpora von korrespondierenden Kriterien (z. B. kurze und lange Dialoge) statt. Konkret wurden die vier Korpuspaarungen (sb, sw), (si, li), (ts, tf) und (w1, w6) verwendet. Training bedeutet hier, dass für die beiden Kopora das N-Gramm-Modell mit den entsprechenden Werten für n und λ erzeugt und dann dem Klassifikator zugeordnet wurden.

Im Rahmen der Leave-One-Out-Kreuzvalidierung erfolgte das Training mit allen Dialogen der Kopruspaarung, außer dem einen Ausgelassenen. Der verbliebene Dialog wurde dann mittels des trainierten Klassifikators einer der beiden trainierten Klassen (z. B. langer Dialog oder kurzer Dialog) zugeordnet. Dies wurde für jeden Dialog in den beiden verwendeten Korpora wiederholt. Für die eigentliche Klassifikation eines Dialogs, wurde zunächst der Abstand des Dialoges zu den beiden trainierten Klassen (Kriterien) mittels des jeweiligen Distanzmaßes berechnet. Der Dialog wurde dann der Klasse, zu der er den geringeren Abstand hatte, zugeordnet. Die Leistung

des Klassifikators wurde mit Hilfe des F-Maßes anhand der Werte für richtig-positive (tp), falsch-positive (fp) und falsch-negative(fn) Zuordnung, summiert über alle Validierungsläufe, bestimmt. Das F-Maß wurde anhand der Formel $F_{tp,fp} = (2 * tp)/(2 * tp + fp + fn)$ ([2, S. 51]) berechnet. Die Ergebnisse der Validierung sind in Tabelle 2 gelistet, welche im Abschnitt 4 weiter erläutert wird.

3.2 Berechnung von Unterschieden zwischen Korpora

Für jede der 96 Kombination aus Distanzmaß, n und λ wurde der der Abstand zwischen den folgenden Dialogkorpora berechnet: (a) kurze und lange Dialoge, (b) erfolgreiche und nicht erfolgreiche Dialoge, Dialoge bei 100 % und 60 % Worterkennungsrate, und schließlich zwischen dem kompletten empirischen Korpus und (d) dem Korpus der guten sowie (e) dem der schlechten Simulation. Eine Auswahl der erhaltenen Resultate ist in Tabelle 3 aufgeführt. Eine weitergehende Betrachtung der Resultate gibt auch hier Abschnitt 4.

4 Resultate

Abschnitt 2 stellt Methoden vor, deren Anwendung auf vorhandene Daten in Abschnitt 3 gezeigt wird. Im Folgenden werden die Resultate, der im vorhergehenden Abschnitt vorgestellten Arbeiten, beschrieben und erläutert. Die Werte in den Tabellen 2 und 3 sind aus Platzgründen nur bis zur zweiten Nachkommastelle angegeben.

Wie oben beschrieben, wurden die Distanzmaße wurden zunächst als Klassifikator verwendet und ihre Klassifikationsleistung mittels Leave-One-Out-Kreuzvalidierung und der Berechnung des $F_{tp,fp}$-Maßes bewertet. Tabelle 2 stellt die $F_{tp,fp}$-Werte für jede der verwendeten Kombinationen einander gegenüber. Die erste Spalte der Tabelle gibt das Kriterium (s. auch Tabelle 1) an, welches der Klassifikator erkennen sollte. Je Kriterium sind die 4 Maße in der zweiten Spalte angegeben. In der ersten Zeile der Tabelle ist der Wert für n bei der Zerlegung der Dialoge in N-Gramme angegeben. Die zu einem Wert für n gehörenden Spalten sind jeweils grau oder weiß hinterlegt. In der zweiten Zeile ist der Wert des verwendeten Faktors bei der Glättung des N-Gramm-Modells (s. Abschnitt 2.2) angeben. Die verwendeten Werte sind: λ_k=0,05, λ_m=0,25 und λ_g=0,5. Bei den leeren Zellen der Tabelle ergab sich für $F_{tp,fp}$ ein Wert von 0. In jeder Zeile ist jeweils der höchste Wert für den jeweiligen Klassifikator (in Abhängigkeit von n und λ) **fett** hervorgehoben. Eine Umrandung markiert pro Kriterium der Dialogauswahl den $F_{tp,fp}$-Wert des besten Klassifikators, also den mit dem höchsten $F_{tp,fp}$-Wert. Sind in dem Block zu einem Kriterium mehrere Werte umrandet, waren diese exakt gleich, auch über die zweite Nachkommastelle hinaus.

Die Distanzen zwischen zwei Korpora wurden, wie in Abschnitt 3.2 beschrieben, mit allen Klassifikatoren berechnet. Tabelle 3 zeigt die Ergebnisse der Distanzberechnung mit der Konfiguration der jeweils besten Klassifikators (Kombination von Distanzmaß sowie Werte für n und λ). Zwischen welchen Korpora die Distanz berechnet wurde gibt jeweils die Spalte P,Q an (s. Tabelle 1). Korrespondierende Werte für die berechnete Distanz (Tabelle 3) und dem $F_{tp,fp}$-Wert des verwendeten Klassifikators sind mit dem gleichen (tiefgestellten) Index versehen. Die Spalten λ und n geben eben diese Werte bei der Berechnung der N-Gramm-Modelle wieder. Die Werte für λ sind wieder: $k = 0,05$, $m = 0,25$ und $g = 0,5$. Für Klassifikatoren, die bei mehreren Konfigurationen den gleichen besten $F_{tp,fp}$-Wert erreichten, sind die Distanzen für all diese Kombinationen angegeben. Die beiden Zeilen mit grauer Textfarbe werden in der Diskussion gesondert besprochen.

n		1			2			3			4			5			6			7			8		
l		λ_k	λ_m	λ_g	λ_k	λ_m	λ_g	λ_k	λ_m	λ_g	λ_k	λ_m	λ_g	λ_k	λ_m	λ_g	λ_k	λ_m	λ_g	λ_k	λ_m	λ_g	λ_k	λ_m	λ_g
ts	c	,58 $_{1}$	,58	,55	,58	,45	,39	,44	,33	,23	,49	,25	,16	,37	,23	,05	,32	,09		,26			,19		
	j	**,54** $_{2}$	,44	,3	,4	,15	,03	,3	,08		,24	,02		,23			,16			,13			,05		
	k	**,56** $_{3}$	,33	,16	,44	,14		,37	,05		,28	,03		,28	,02		,28			,26			,22		
	m	**,56** $_{4}$	,42	,24	,4	,15	,03	,3	,08		,25			,24			,2			,17			,09		
tf	c	,5	,5	,5	,52	,52	,54	,53	,57	,57	,52	,56	,55	,42	**,57** $_{5}$	,55	,47	,56	,54	,44	,54	,54	,45	,54	,54
	j	,55	,57	**,57** $_{6}$	,55	,56	,54	,55	,55	,54	,56	,54	,54	,57	,54	,54	,56	,54	,54	,54	,54	,54	,53	,54	,54
	k	,55	**,58** $_{7}$	,55	,54	,56	,54	,53	,54	,54	,51	,54	,54	,51	,54	,54	,51	,54	,54	,46	,54	,54	,46	,54	,54
	m	,55	,59 $_{8}$	,56	,53	,55	,54	,55	,55	,54	,55	,54	,54	,56	,54	,54	,56	,54	,54	,53	,54	,54	,52	,54	,54
li	c	,98	,97	,97	,99	1 $_{9}$	1 $_{10}$	,88	,88	,91	,5	,57	,75	,32	,56	,78	,25	,59	,83	,2	,6	,77	,25	,63	,72
	j	,98	**,99** $_{11}$	,97	,98	,89	,63	,83	,59	,41	,27	,41	,39	,21	,52	,57	,24	,74	,83	,27	,75	,74	,35	,71	,67
	k	**,99** $_{12}$	,98	,91	,98	,9	,53	,89	,61	,33	,3	,39	,31	,28	,46	,53	,22	,66	,78	,27	,71	,74	,3	,71	,67
	m	,96	**,98** $_{13}$	,95	,94	,89	,59	,81	,61	,44	,29	,41	,39	,25	,5	,56	,22	,7	,82	,27	,73	,73	,34	,7	,67
si	c	,98	,97	,97	,99	1 $_{9}$	1 $_{10}$	,9	,9	,92	,74	,77	,83	,69	,76	,83	,64	,76	,85	,57	,67	,66	,56	,66	,51
	j	,98	**,99** $_{11}$	,97	,98	,91	,78	,86	,77	,73	,69	,73	,72	,61	,74	,77	,55	,76	,79	,33	,59	,42	,1	,28	
	k	**,99** $_{12}$	,98	,92	,98	,92	,75	,9	,78	,71	,59	,72	,71	,53	,71	,75	,45	,67	,68	,3	,46	,42	,07	,28	
	m	,96	**,98** $_{13}$	,95	,94	,91	,77	,86	,78	,73	,67	,73	,72	,56	,74	,76	,54	,74	,8	,31	,57	,41	,12	,28	
w1	c	,57 $_{14}$	,57 $_{15}$	,57 $_{16}$	,57 $_{17}$	,53	,51	,48	,33	,26	,48	,27	,26	,35	,26	,26	,24	,26	,26	,22	,26	,26	,24	,26	,26
	j	,57 $_{18}$	,47	,35	,47	,28	,26	,35	,27	,26	,3	,26	,26	,28	,26	,26	,26	,26	,26	,26	,26	,26	,25	,26	,26
	k	**,48** $_{19}$	,31	,26	,47	,28	,26	,35	,27	,26	,31	,26	,26	,3	,26	,26	,26	,26	,26	,24	,26	,26	,23	,26	,26
	m	**,51** $_{20}$	,42	,3	,51	,32	,26	,34	,27	,26	,31	,26	,26	,3	,26	,26	,27	,26	,26	,26	,26	,26	,25	,26	,26
w6	c	,89	,89	,89	,91 $_{17}$	,87	,83	,87	,62	,17	,87	,17		,76	,02		,52			,28			,26		
	j	**,89** $_{18}$	,78	,57	,81	,28	,04	,62	,11		,34			,23			,04			,06			,06		
	k	**,86** $_{19}$	,51	,15	,81	,28	,04	,69	,17		,51	,02		,39	,02		,29			,21			,16		
	m	**,88** $_{20}$	,71	,41	,83	,42	,04	,67	,13		,39	,02		,31			,13			,06			,06		
sb	c	.57	.56	.52	.58	.6	.66	.63	.69	.69	.64	.69 $_{21}$	.67	.61	.69	.67	.58	.68	.67	.58	.67	.67	.57	.65	.67
	j	.5	.29	.18	.42	.36	.34	.44	.39	.39	.51	.67	**.68** $_{22}$	.63	.68	.67	.68	.67	.67	.68	.67	.67	.67	.67	.67
	k	.3			.34	.06	.01	.36	.1	.01	.38	.2	.03	.4	.26	.11	.54	**.68** $_{23}$	.68	.65	.67	.67	.66	.67	.67
	m	.42	.15	.03	.41	.3	.26	.45	.4	.39	.64	**.69** $_{24}$	.68	.68	.67	.67	.68	.67	.67	.67	.67	.67	.67	.67	.67
sw	c	.63	.64	**.66** $_{25}$	.65	.64	.56	.6	.53	.47	.52	.52	.07	.48	.38		.46	.37		.46	.37		.46	.46	.01
	j	.66	.7	.69	.7	.72	.71	.71	.72	.72 $_{26}$	.69	.49	.2	.58	.09		.32			.13			.06		
	k	.69	.67	.67	**.71** $_{27}$	.67	.67	.71	.68	.67	.71	.69	.67	.69	.7	.68	.63	.4	.13	.35			.18		
	m	.67	.68	.67	.7	.71	.7	.71	.71	**.72** $_{28}$	.58	.45	.11	.47	.01		.28			.09			.05		

Tabelle 2 - Gegenüberstellung der Werte des $F_{tp,fp}$-Maßes der Klassifikatoren unter verschiedenen Bedingungen. Abkürzungen in Spalte 2: c = Kosinusmaß, j = Jensen Differenzen Distanz, k = Kullback-Leibler Divergenz, m = Mittlere Kullback-Leibler Divergenz.

Kosinus				Jensen				Kullback-Leibler				Mittlere K.-L.			
P, Q	λ	n	D	P, Q	λ	n	D	P, Q	λ	n	D	P, Q	λ	n	D
ts, tf	k	1	$,05_{1}$	ts, tf	k	1	$,03_{2}$	ts, tf	k	1	$,14_{3}$	ts, tf	k	1	$,14_{4}$
ts, tf	m	5	$,12_{5}$	ts, tf	g	1	$,03_{6}$	ts, tf	m	1	$,12_{7}$	ts, tf	m	1	$,12_{8}$
li, si	m	2	$,69_{9}$	li, si	m	1	$,25_{11}$	li, si	k	1	$1,49_{12}$	li, si	m	1	$1,29_{13}$
li, si	g	2	$,68_{10}$												
w1, w6	k	1	$,15_{14}$	w1, w6	k	1	$,062_{18}$	w1, w6	k	1	$,29_{19}$	w1, w6	k	1	$,29_{20}$
w1, w6	m	1	$,14_{15}$												
w1, w6	g	1	$,14_{16}$												
w1, w6	k	2	$,27_{17}$												
ex, sb	m	4	$,8_{21}$	ex, sb	g	4	$,46_{22}$	ex, sb	m	6	$3,19_{23}$	ex, sb	m	4	$4,04_{24}$
ex, sw	m	4	,85	ex, sw	g	4	,47	ex, sw	m	6	3,38	ex, sw	m	4	4,22
ex, sw	g	1	$,57_{25}$	ex, sw	g	3	$,52_{26}$	ex, sw	k	2	$6,32_{27}$	ex, sw	g	3	$3,99_{28}$
ex, sb	g	1	,54	ex, sb	g	3	,49	ex, sb	k	2	5,79	ex, sb	g	3	3,69

Tabelle 3 - Distanzen zwischen zwei Korpora in Abhängigkeit vom verwendeten Distanzmaß und den Werten für n und λ. Abkürzungen: $k = 0,05$, $m = 0,25$ und $g = 0,5$. Neben den Kürzeln aus Tabelle 1, steht ex für das komplette empirischen Korpus.

5 Diskussion und Ausblick

Einige Eigenschafen der verwendeten Maße müssen hier kritisch betrachtet werden. Die Kullback-Leibler Divergenz ist nicht symmetrisch, weshalb an ihrer Stelle immer die Mittlere KLD verwendet werden sollte, wenn sie als Distanzmaß eingesetzt wird. Weiterhin haben KLD-basierte Maße keine obere Grenze, weshalb sie nur verwendet werden können, um den Abstand von 2 Korpora zu einem Dritten zu vergleichen. Eine Abschätzung wie (un)ähnlich ein einzelner Korpus im Vergleich zu einem Anderen ist kann aber nicht vorgenommen werden.

Die Jensen Differenzen Distanz und die Kosinus Distanz liegen immer zwischen 1 und 0, allerdings ist die Kosinusdistanz anschaulicher in ihrer Anwendung und hat sich auch bei anderen N-Gramm-basierten Anwendungen bewährt.

Bei der Anwendung auf acht verschiedene Kriterien ist ein Klassifikator der den Kosinunsabstand verwendet 6 mal der Beste, 2 mal einer auf Basis der Jensen Differenzen Distanz und einmal einer der die Mittlere Kullback-Leibler Divergenz verwendet (wobei für Kosinusdistanz und JD in einem Fall die $F_{tp,fp}$-Wert identisch waren).

Bei der Betrachtung der Distanzen in Tablle 3 fällt auf, dass die Distanz zwischen den Korpora mit erfolgreichen (ts) und nicht erfolgreichen (tf) Dialogen für jedes Maß sehr gering ist. Der Grund dafür liegt vermutlich darin, dass sich diese Dialoge oft nur im letzten Dialogakt unterscheiden, was entsprechend ähnliche N-Gramm-Modelle zur Folge hat.

In Tabelle 2 zeigt sich, dass die Werte für n und λ durchaus einen Einfluss auf die Leistung eines Klassifikators haben. Besonders auffällig ist der Unterschied des optimalen Wertes für n, bei Klassifikatoren die auf empirischen Daten arbeiten und solchen die mit Simulationsdaten operieren. Deshalb ist die Anwendung der in Abschnitt 3.1 beschriebenen Methode zur Bestimmung der optimalen Konfiguration sinnvoll. Um den Aufwand an dieser Stelle zu reduzieren kann ausschließlich der Kosinusabstand als Distanzmaß verwendet werden, da dieser – bei gut gewähltem n und λ – zu einer guten Klassifikatorleistung und somit auch einer verlässlichen Distanzberechnung führt.

In Tabelle 3 enthalten die Zeilen mit grauem Text nicht die Distanzen die mit einer optimalen Konfiguration ermittelt wurden. Vielmehr wurde die Konfiguration aus der jeweils darüber

stehende Zeile verwendet, um den Abstand zwischen den empirischen Dialogen und denen der jeweils anderen Simulation zu berechnen. Das Ergebnis zeigt, dass alle Distanzmaße einen größeren Unterschied zwischen der schlechten Simulation und dem empirischen Korpus ermitteln, als zwischen der guten Simulation und den empirischen Dialogen. Dieses Resultat entspricht auch der Erwartung, da die bessere Simulation auch realistischere Dialoge erzeugen soll.

In den sich nun anschließenden Arbeiten soll u. a. untersucht werden ob und wie sich bestimmen lässt, welche der Klassifikatorkonfigurationen sich besonders gut zur Messung des Abstandes zwischen empirischen und simulierten Dialogkorpora eignet. Für diese Entscheidung könnten unter anderem Klassifikatoren herangezogen werden, die darauf trainiert sind zwischen simulierten und empirischen Dialogen zu unterscheiden.

Literatur

[1] BIGI, B.: *Using Kullback-Leibler Distance for Text Categorization*. In: SEBASTIANI, F. (Hrsg.): *Advances in Information Retrieval*, Bd. 2633 d. Reihe *Lecture Notes in Computer Science*, S. 305–219. Springer Berlin Heidelberg.

[2] FORMAN, G. und M. SCHOLZ: *Apples-toApples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement*. SIGKDD Explorations, 12(1):49–57, Juni 2010.

[3] GEORGILA, K., J. HENDERSON und O. LEMON: *User Simulation for Spoken Dialogue Systems: Learning and Evaluation*. In: *Proc. 9th Int. Conf. Spoken Lang. Process.*, S. 1065–1068, Pittsburgh, USA, 2006.

[4] HILLMANN, S. und K.-P. ENGELBRECHT: *Aufgabenmodellierung in der Simulation von Interaktionen mit Sprachdialogsystemen*. In: WAGNER, P. (Hrsg.): *ESSV 2013*, Bd. 65 d. Reihe *Studientexte zur Sprachkommunikation*, S. 20–27. TUDpress, März 2013.

[5] KULLBACK, S. und R. A. LEIBLER: *On Information and Sufficiency*. The Annals of Mathematical Statistics, 22(1):79–86, 1951.

[6] MÖLLER, S.: *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, New York, United States, 2005.

[7] PIETQUIN, O. und H. HASTIE: *A survey on metrics for the evaluation of user simulations*. The Knowledge Engineering Review, 28(1):59–73, 2013.

[8] SALTON, G. und C. BUCKLEY: *Term-weighting approaches in automatic text retrieval*. Information Processing & Management, 24(5):513–523, 1988.

[9] TANEJA, I. J.: *Generalized Information Measures and Their Applications*. 2001.

[10] WILLIAMS, J. D.: *Evaluating user simulations with the Cramér-von Mises divergence*. Speech Communication, 50(10):829–846, Okt. 2008.