Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Mensch-Technik-Interaktion
SmartKom

Gestaltung der End–to–End–Evaluation in SmartKom

Version 2.0

Nicole Beringer, Ute Kartal,
Marion Libossek, Silke Steininger

Ludwig–Maximilians–Universität München

Technisches Dokument Nr. 19

April 2002

April 2002

Nicole Beringer
Ute Kartal
Marion Libossekl
Silke Steininger

IPSK LMU München
Schellingstr. 3
80799 München

Tel.: (089) 2180−5751 oder 5753
FAX: (089) 2800362

E−Mail:{beringer|ukartal|kstein}@phonetik.uni−muenchen.de

Dieses Technische Dokument gehört zu Projektfeld II: Anwendungsszenarien

Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

Inhaltsverzeichnis

1 Einleitung

Als End–to–End–Evaluation definieren wir die »Evaluation der Aufnahme–Prototypen SmartKom Public / Mobile / Home mittels Benutzerexperimenten.« (lt. Projektantrag [SK–Antrag]).

Aufgabe der End–to–End–Evaluation ist die Untersuchung der Benutzbarkeit, Effizienz und Akzeptanz der zugrunde liegenden **Konzepte** der Prototypen SmartKom–P / M / H für die multimodale Kommunikation parallel zur Datenerhebung.

Unter Konzept ist hier der Einsatz und die Kombination verschiedener Modalitäten für die Ein– und Ausgabe, nicht jedoch die aktuelle Ausarbeitung der einzelnen Anwendungsfälle zu verstehen.

Neben den WOZ–Aufnahmen sollen für die einzelnen Szenarien Evaluationen anlaufen. Die Ergebnisse – besonders von Benutzbarkeit und Effizienz – werden, da sie im Entwicklungszeitraum der Prototypen entstehen, sofort wieder an die Partner weitergegeben. Somit ist gewährleistet, daß die Module einen unmittelbaren Nutzen daraus ziehen können.

Die Parallelität der End–to–End–Aufnahmen zu den normalen Wizardaufnahmen soll zu einer Weiterentwicklung der Prototypen aufgrund der Erkenntnisse der Evaluationsexperimente beitragen.

Entgegen anfänglicher Planung werden keine Module durch Wizards ersetzt. An einigen wenigen Stellen werden Funktionalitäten simuliert (z.B. wird nicht tatsächlich ein Fax aufgenommen). Abgesehen von solchen kleinen Abweichungen von einer tatsächlich echten Anwendung wird immer der gesamte Demonstrator evaluiert, möglichst mit allen physisch vorhandenen Peripheriegeräten.

2 Begriffsklärung

2.1 Benutzbarkeit

Darunter verstehen wir (im Gegensatz zu dem Fachbegriff aus dem Usability Engineering) das Funktionieren bzw Nicht–Funktionieren der Module und deren Koordination. Herauszufinden gilt, ob der Benutzer alle Funktionen des Systems aus technischer Sicht nutzen kann.
Methode:
•indirekt durch standardisierte Experimente,
•direkt durch »Trockentest« ohne VP (isoliert bzw. in Zusammenarbeit (Schnittstellen))
(s. hierzu Technische Evaluation).

2.2 Effizienz

Die Betrachtung der Effizienz teilt sich auf in:

a) objektive Effizienz:
•Wie leicht / schwer ist das System für den Benutzer zu bedienen? Also, wieviele **Fehler** / Nachfragen etc. treten bei der Bearbeitung von definierten Standardaufgaben auf?
•Wie verhält sich das System in Bezug auf **Reaktionszeiten** bei der Bearbeitung der Anfrage / Ausgabe?
Interessant könnte hier auch andere objektive Bewertungskriterien des Umgangs des Benutzers mit dem System sein, also z.B. wieviel Off–Talk tritt auf, wenn das System auch auf Äußerungen des Benutzers reagiert, die nicht an das System gerichtet waren oder auch die Reaktion auf Verhalten des Systems bei Barge–In.
Die Auswertung findet im Rahmen standardisierter Experimente statt unter

Berücksichtigung der analysierten System−Protokolle der Aufnahmen.

b) subjektive Effizienz:
•Wie bewertet der Benutzer (über Interview[1] / Fragebogen) die Arbeit mit dem System insgesamt?
•Welchen Eindruck hinterläßt die Dateneingabe / Benutzerführung / Datenausgabe?
•Wie werden Wartezeiten / Fehler / Nachfragen aus der Sicht des Benutzers bewertet?

Methoden: Fragebogen, Interview, indirekte Methoden.

Der Fragebogen wird dynamisch zusammengestellt, je nachdem, welche Anwendungen in dem jeweiligen Szenario vorkommen.
Es werden parallele Tests für die objektive und subjektive Effizienz durchgeführt und danach unter den Aspekten:
•Tatsächliche Leistung des Systems (technisch orientiert) im Gegensatz zur
•empfundenen Leistung (aus ergonomischer Sicht)
ausgewertet.

2.3 Akzeptanz

Darunter verstehen wir die Einstellung der Versuchspersonen gegenüber dem System, NACHDEM sie das System verwendet haben (man könnte hier ja auch zusätzlich davor nachfragen). Nur einige wenige Einstellungen werden bereits vor der Benutzung des Systems erfragt (z.B. Zugang/Nutzung von Computern allgemein, Selbsteinschätzung Computer−Kompetenz, Einstellung zu Emotionsaufzeichnung und −analyse) .
Methoden: Fragebogen, Interview, indirekte Methoden.

Interviews oder indirekte Methoden können dazu dienen, Informationen zu erhalten, die nicht explizit erfragt werden können (weil sie nicht bewußt sind oder weil sie aus bestimmten Gründen nicht/wahrscheinlich nicht geäußert werden). Weil sie sehr zeitintensiv sind, finden sie nicht bei allen, sondern nur einer ausgewählten kleinen Gruppe von Benutzern statt.

2.4 Konzepte

»Zielstellung:
Im Rahmen dieses Arbeitspakets soll untersucht werden, inwieweit die zugrunde liegenden Konzepte − realisiert in Aufnahme−Szenarien − des Prototypen SmartKom− von Benutzern aus unterschiedlichen Benutzergruppen (naiv bis Experte) akzeptiert werden (subjektive Evaluation mittels Interview). Ergebnisse sollen den Design−Prozess positiv beeinflussen.«

Dabei gilt zu bedenken:

•Wir können nur vorhandene Module betrachten und diese nur, wenn sie technisch funktionsfähig sind.
•Außerdem läßt sich das System nur als Ganzes betrachten, weil die Evaluation von einzelnen Modulen nicht in den Bereich der End−to−End Evaluation fällt.
•Wird das System durch das Verhalten eines Moduls »unbenutzbar« (extrem lange Wartezeiten, Fehler z. B. in mehr als 3 / 4 aller Anfragen / Ausgaben kann eine ergonomische Evaluation nur eingeschränkt stattfinden. Nicht abfragbare Kriterien müssen in diesem Fall von Experten über eine Checkliste bewertet werden − was

[1] Interviews finden nur bei einem kleinen Teil der Benutzer statt, der Großteil beantwortet einen standardisierten Fragebogen.

natürlich um ein Vielfaches weniger valide Ergebnisse liefert.

3 Allgemeine Kriterien einer End–to–End Evaluation von Dialogsystemen

Die Performanz einer Evaluation wird durch Charakteristika des zu bewertenden Systems beeinflußt [Andenfilger–97]. Die Hauptaspekte sind folgende:

- Die Bedürfnisse der Entwickler
  Eine End–to–End Evaluation muß maßgeblich auf die Qualität der Mensch–Maschine–Interaktion ausgerichtet sein. Dabei muß eine Bewertung der Performanz des Systems unter realistischen Anwendungsbedingungen erfolgen, um allgemeine oder teilweise Verbesserungen des Systems feststellen zu können. Im Fall SmartKom muß ein besonderes Augenmerk auf der Korrelation der Eingabe– und Ausgabemodalitäten liegen.
- Die Anforderungen der Benutzer
  Durch die Befragung der Benutzer werden die Anforderungen an die Funktionalitäten eines multimodalen Dialogsystems untersucht.
- Die Einschränkungen der Evaluation von Dialogsystemen bzw. multimodalen Systemen im allgemeinen:
  Die hohe Variabilität von Eingabe und Ausgabemöglichkeiten sowie die schlechte Vergleichbarkeit mit anderen Systemen.

Es wird eine allgemeine Bewertung des Systems durch den direkten Vergleich von ergonomischen Evaluationsgesichtspunkten mit objektiv meßbaren Werten des jeweiligen Testlaufs zu erreichen versucht (s. hierzu PROMISE).

4 Zeitliche und inhaltliche Aufteilung der Evaluation

»Dabei sollen neben technischen Aspekten (Funktionalität, Effizienz) auch ergonomische Gesichtspunkte (Bedienbarkeit, intuitive Bedienung, etc.) untersucht werden.«

Die End–to–End–Evaluation wird auf die Bereiche „Teschnische Evalutation“ und „Ergonomische Evaluation“ aufgeteilt, siehe folgende Abschnitte.
Als Grundlage der Evaluation wird das Testbed mit den zu verschiedenen Zeiten verfügbaren Modulen verwendet.

4.1 Technische Evaluation (TE): Benutzbarkeit, objektive Effizienz

Bei der technischen Evaluation wird die Funktionalität des Systems in ihren Einzelheiten (s. Kapitel 5) untersucht. Ist das System funktionstauglich, wenn nicht, bei welchen Modulen gibt es Probleme? (Benutzbarkeit) Daneben werden objektiv bewertbare technische Daten des Systems (Wie schnell, wie viele Fehler, Echtzeitfaktor bei der Spracherkennung usw.) ausgewertet (objektive Effizienz) bzw. zur Bewertung an die ergonomische Evaluation weitergegeben.
Die TE teilt sich wiederum in 3 Bereiche:

1. »Trockentest«
   Für die Effizienz der End–to–End–Evaluation ist es wichtig, daß die jeweiligen Endergebnisse aller Module bewertet werden. Das soll allerdings nicht heißen, daß die einzelnen Module auf spezifische Fehler hin analysiert werden, sondern, daß die Kommunikationshäufigkeit bzw. –erfolge (nur »hat funktioniert«, »hat nicht funktioniert, weil Schleife / tot«) analysiert wird. (Diese Funktionalität wird bei der Installation des

Systems berücksichtigt. Gefundene Probleme, Bugs etc. werden an sk−moko bzw. an die sk−Systemgruppe zurückgemeldet. Es wird keine weitere Veröffentlichung der Ergebnisse durchgeführt.) Der „Trockentest“ dient nur zur internen Problemlösung. Wir erwarten ein lauffähiges System mit funktionierenden Modulen. Die LMU führt eine End−to−End−Evaluation, kein Beta−Testing durch!

2. Evaluation der Schnittstelle Mensch − System
   Das betrifft alle Module oberhalb der Linie Interpretation/Analyse − Generierung (s. Graphik). Sie werden bezüglich ihrer Benutzbarkeit und Effizienz analysiert.
3. Standardisierter Vergleich der subjektiven Ergebnisse der Benutzerbefragung mit den objektiv meßbaren Daten aus der TE durch die Verwendung eines graphischen Evaluationstools (S. 9.3).

4.2 Ergonomische Evaluation (EE): subjektive Effizienz, Akzeptanz

Die ergonomische Evaluation bewertet den subjektiven Bereich der Effizienz, sowie die Akzeptanz mit Hilfe von Benutzerbefragung mit Fragebogen und Interview
Die Benutzerevaluation finden mithilfe eines standardisierten online−Fragebogens statt. Einige Fragen werden der Versuchsperson vor Beginn des Experimentes gestellt, der restliche Fragebogen wird nach erfolgtem Test beantwortet. Dieser Fragebogen enthält Fragen zu der subjektiven Effizienz sowie Akzeptanz des Systems SmartKom.

- Subjektive Effizienz:
  Wie bewertet die VP die Effektivität des Gerätes zur Lösung von bestimmten Problemen insgesamt?
  Wie werden die Eingabemedien bzw. Ausgabemedien, sowie der Dialog MM allgemein bewertet?
  Wie bewerten die VP die Leistung speziell für den durchgeführten Fall? (Anwendungen, Szenario)
- Welche subjektiven Schwierigkeiten traten bei der Bedienung auf. Kommunikationsmittel ausreichend?
- Akzeptanz:
  Was ist die Einstellung der VP dem System gegenüber allgemein?
  Wie werden die Eingabemedien bzw die Ausgabemedien, sowie die verschiednen Methoden der Biometrie empfunden?

  Zu welchen Personen würde dieses System passen, bzw. wie sehen die VP das System (z. B. durch ein Übertragen auf Fahrzeugtypen)
- Allgemeines Wohlbefinden bei der Nutzung von SmartKom, Attraktivität/Akzeptanz einzelner Komponenten.

  Um allen Anforderungen gerecht zu werden, wird die End−to−End Evaluation in zwei Zeitbereiche aufgeteilt.

4.3 Begleitende Evaluation − Möglichkeit zur Verbesserung

Durch die Aufteilung der Evaluation in zwei Zeitbereiche pro Szenario erhalten die Entwickler die Möglichkeit, unter Verwendung aller Ergebnisse und Protokolle inhaltliche und technische Probleme zu lösen. Dafür werden von seiten der Evaluation alle objektiven und subjektiven Ergebnisse zur Verfügung gestellt; ein direkter Vergleich erfolgt durch das in Abschnitt 9.3 beschriebene graphische Evaluationstool, mit dem die wichtigsten Gegenüberstellungen abgerufen werden können.

Durch diesen Service kann das System zielgerichtet für die eigentliche Evaluation – die abschließende Evaluation verbessert werden.

4.4 Abschließende Evaluation – Eigentliche Bewertung des Systems

In der zweiten Phase der Evaluation wird jedes Szenario unter denselben technischen Voraussetzungen bewertet. D.h. mit Beginn der ersten Evaluationsläufe werden in allen Szenarien keine Verbesserungen mehr akzeptiert. Alle Ergebnisse werden dem Projekt nach Beendigung einer Szenarioevaluation zur Verfügung gestellt.
Anhang A enthält den Zeitplan für die einzelnen Evaluationszeiträume.

5 Evaluationskriterien

5.1 Paradise: A Framework for Evaluating Spoken Dialogue Agents

5.1.1 Entkoppelung von Anforderungen aus dem Anwendungsfall und dem Avartarverhalten (dem Dialogverhalten der Maschine)

Die folgenden Abschnitte geben eine Zusammenfassung aus [1], [2], [3], [4], [5], [6], [7] über die Funktionalität und die Verarbeitungsweise von PARADISE (PARAdigm for DIalogue System Evaluation (von DARPA Communicators).

Bisher:
Evaluation der Dialogstrategien über „inappropriate utterance ratio, turn correction ratio, concept accuracy, implicit recovery, transaction success“: Robustheit des Agenten (Systems) ergibt sich über die Wertung der Teilergebnisse verschiedener Dialogbewertungen.
(Dialogstrategie: Wie wird ein Dialog vom System geplant? Z.B. systemgeführter oder usergeführter Dialog, gemischter Dialog, vom allgemeinen zum besonderen, Komplettvorschläge)

Evaluation mit vordefinierter Referenzantwort [7]: Die Systemantwort auf eine Benutzerfrage wird mit einem vordefinierten Schlüssel von minimalen und maximalen Referenzantworten verglichen. Die Performanz ist der Anteil der Antworten, die auf den Schlüssel zutreffen. Nachteil: Es wird nur eine Dialogstrategie zugelassen.

Beide Ansätze können nicht auf multimodale Dialogsysteme generalisiert werden. Beide Ansätze enthalten zudem Faktoren, die Einfluß auf die Berechnung der Performanz nehmen, jedoch nicht im Detail nachvollziehbar sind. Auch Faktoren wie Datengröße und Umweltfaktoren wie Hintergrundgeräusche beeinflussen die Performanz.

Lösung: PARADISE
1. Vergleich von Dialogstrategien durch die Entkoppelung der benötigten Anwendungsfall–Anforderungen von der Art der Bearbeitung via Dialog. Die Performanz ergibt sich als eine gewichtete Funktion des anwendungsfallbasierten Erfolgsmaßes und dialogbasierter Kostenmessung, wobei sich die Gewichte aus der Korrelation von Benutzerzufriedenheit mit der Performanz ergeben.
2. Evaluierung des Prozentsatzes korrekter Antworten bzgl. einer Referenzantwort, der Prozentsatz erfolgreicher Transaktionen oder vervollständigter Tasks, Prozentsatz von angemessenen und unangemessenen systemdirektiven und diagnostischen Äußerungen, User Satisfaction (über Attribute Value Matrix).

3. Messungen können sehr stark korrelliert sein und das wird bei anderen Evaluationen nicht berücksichtigt.
4. Paradise erlaubt Generalisierungen zwischen Messungen.

5.1.2 Vergleich von Dialogstrategien

Performanzmodell für Dialoge: Methoden aus der Entscheidungstheorie für die Kombination von einer ungleichartigen Menge von Performanzmeßwerten (user satisfaction, task success, dialogue cost) in eine einzige Performanzfunktion zur Evaluation (s. Anhang C) unter Verwendung einer

ATTRIBUTE VALUE MATRIX.
Eine Attribute Value Matrix (AVM) stellt verschiedene Tasks des Dialogs dar.
Die Informationen, die zwischen System und Benutzer ausgetauscht werden, werden als eine Reihe von geordneten Attribut−Paaren und deren möglichen Werten dargestellt.

Der Erfolg einer Task (nicht dynamischer Anwendungsfall wie in SK sondern statisch (abstrakte) Aufgabenstellung) für einen ganzen Dialog (oder Subdialog) wird dadurch gemessen, wie gut das System und der Benutzer die Informationsanforderungen eines bestimmten Tasks erreichen. Dieselbe Task wird in mehreren Dialogen durchgeführt. Die Anwendungsfälle können dabei variieren.
Hieraus wird durch multiple lineare Regression (MLR) eines z−normalisierten (s. Anhang C) Kappa−Koeffizienten und der User Satisfaction der Task Success berechnet.

Daneben werden einzelne Kostenfunktionen (qualitative und quantitative objektive Messungen) für den Dialog berechnet. Diese werden normalisiert und ebenfalls mit der gesamten User Satisfaction über MLR korreliert.

5.1.3 Generalisierung

Sinn des PARADISE −Frameworks ist eine Generalisierung von Dialogstrategien und Systemverhalten wie folgt:

- Über die AVM kann jede Task, jede Domäne und jeder Dialog abgearbeitet werden, allerdings wird nicht unterschieden, wann falsche und wann nicht vorhandene Teiltasks der Task Success schmälern.
- Die Attribute müssen allerdings angepaßt werden.
- Der Dialog muß respektive in seine Dialogakte aufgeteilt werden, auf die dann die Attribute gesetzt werden.
- Genau festgelegte Arbeitsschritte
- Unterstützung von Vergleichen zwischen Dialogstrategien,Task−Darstellung bei verschiedenen Dialogsystemen.
- Task success measure: Normalisiert für Taskkomplexität =>Basis für den Vergleich verschiedener Tasks.
- Performanzevaluation auf jeder Dialogebene.
- Kombination von Erfolgsmaßen (objektive und subjektive Kosten), <u>Spezifizierung der relativen Beiträge quantifiziert auf gesamte Performanz. Nicht ganz klar</u>
- Integration von Transaction success, Konzeptakkuratheit, Kostenfunktionen, User Satisfaction bei der Berechnung der Performanz verschiedener Systeme.
- Die Performanz der Systeme ergibt sich als eine lineare Kombination von Task Success und Dialogkomplexität

6 Probleme bei der Evaluation mit PARADISE

6.1 Ergonomische Evaluation

Viele in PARADISE verwendete Bewertungskriterien können nicht oder nur eingeschränkt abgefragt werden, da SmartKom nicht mit unimodalen Dialogsystemen verglichen werden kann.
Daher müssen einige Fragen aus dem Paradise−Fragenkatalog komplett weggelassen bzw. ersetzt werden. Das sind im wesentlichen folgende:
„Comparable Interface“: Es gibt keine multimodalen Standardsysteme, mit denen verglichen werden kann.
„Interaction Pace (Bewertung des Gesamtdialogs) vs. System Response (Bewertung der Systemantworten)“: Wie kann eine VP objektiv einen Dialog auswerten, bei dem sie selbst mitwirkt?
Die Skalen zu den einzelnen Fragen sind nicht homogen (3 vs. 5 Wertungen) und zudem nicht auf die einzelnen Kostenfunktionen zuweisbar.

6.2 Technische Evaluation

Während „Interaction Pace vs. System Response“ bei der Benutzerbefragung dieselbe Antwort ergeben, kann bei der TE eine objektiv meßbare Unterscheidung zwischen Dauer des Gesamtdialogs und der mittleren Dauer der Systemantwortzeit getroffen werden. Allerdings kommt es auch hier zu einigen grundsätzlichen Problemen.

6.3Probleme bei der Portierung der Evaluationsrichtlinien von PARADISE auf multimodale Dialogsysteme

Die Probleme, die sich bei der Portierung der Evaluationsrichtlinien von PARADISE auf SmartKom ergeben, sind offensichtlich:
1. Wie geht man mit multimodalen Eingabe− und Ausgabemedien um?
2. Wie verhält es sich mit der Evaluation der drei Szenarien? Wird hier ebenso wie über die Anwendungsfälle normalisiert? Wenn ja, wie? Wenn nein, haben wir dann eigentlich 3 Systeme?
3. Bei der Berechnung der AVMs für SmartKom entstehen multidimensionale Matrizen. Wie kann man hier einen optimalen Dialog rausfinden?
4.Wie geht man mit der so entstehenden Streuung um?

Die Lösung der LMU wird nachfolgend zusammengefaßt.

7 PROMISE − PROcedure for Multimodal Interactive System Evaluation

Die publizierte Form von PARADISE ist für SmartKom nicht anwendbar =>

Die Evaluation multimodaler Systeme muß klar von PARADISE unterschieden werden.

7.1 Einschränkungen multimodaler Dialogsysteme und deren Evaluation

- Unterscheidung der Natürlichkeit der Eingabe und der Natürlichkeit der Ausgabe: Die Verbesserung der Sprach−, Gestik− und Mimikerkenner in den letzten Jahren erlauben eine natürliche Eingabe durch den Benutzer.
- Die Ausgabe ist trotz der Verbesserung von Sprachsynthese und graphischen Darstellungsmodalitäten in ihrer Natürlichkeit vom Zusammenspiel mehrerer

Faktoren abhängig:
In der Wahl der Darstellung des Agenten (meist als Cartoon) müssen die Entwickler sowohl darauf achten, daß dieser einerseits virtuell aber doch natürlich, andererseits kompetent und doch nicht unbedingt menschlich aber in jedem Fall multimodal ist. Z.B. muß die Synthesestimme zur graphischen Darstellung passen, also z.B. eine Komikfigur mit einer eher kindlichen Stimme, wenn auch diese aus synthesetechnischer Sicht eher ungeeignet ist.

- Abstraktion vom eigentlichen System
  Die wichtigste Einschränkung bei der Evaluation multimodaler Systeme ist die schwierige Vergleichbarkeit mit anderen Dialogsystemen bzw. –szenarien, sowie die Abstraktion von Dialogstrategien. Diese Probleme werden bei SmartKom noch verstärkt durch das Fehlen von statischen Anwendungen – zweifellos ein Plus für das System, jedoch sind dynamisch definierte Anwendungen nicht einfach miteinander vergleichbar. Der Ansatz der Berechnung einer Task Success durch Referenzschlüssel und Attribute Value Matrizen [1, Hirschmann−90] würde im Fall SmartKom−EPG eine zweidimensionale Matrix mit meheren tausend Reihen und Spalten ergeben, deren Werte alle möglich werden. Die Wahrscheinlichkeit einer Abweichung vom Wert und vom Schlüssel ist wesentlich höher und nicht unbedingt ein Anzeichen eines Fehlers.

7.2 Funktionserweiterungen bzw. –änderungen bzgl. PARADISE

Aufgrund der oben beschriebenen Probleme ist SmartKom evaluiert mit Paradise nicht eins zu eins vergleichbar mit der Paradiseevaluation von Spoken Dialogue Systems.
Neben den bereits vorhandenen Funktionen aus Paradise müssen multimodale Funktionen erweitert (z.B. Dialogdauer) bzw. in Anlehnung an den Sprachbereich neu definiert werden.

Nr	**FUNKTION bei SmartKom**	**Nomenklatur nach Paradise**
1	Dialogdauer	Dialog elapsed time
2	Fehlertoleranz bei Anfragen	Fehlertoleranz bei Anfragen (cooperativity)
3	Fehlerfrequenz von Anfragen	Fehlerfrequenz von Anfragen (timeout / timeout %)
4	Barge−In	Bargein, bargein %
5	Effizienzwerte	Performance
6	Dauer der „Sprecheingabe“	Elapsed time der Sprecheingabe
7	Dauer der Spracherkennung	Elapsed time der Spracherkennung

Nr	**FUNKTION bei SmartKom**	**Nomenklatur nach Paradise**
8	Dauer der „Mimikeingabe“	Elapsed time der Mimikeingabe
9	Dauer der „Mimikerkennung“	Elapsed time der Mimikerkennung
10	Dauer der „Gestikeingabe“	Elapsed time der Gestikeingabe
11	Dauer der „Gestikerkennung“	Elapsed time der Gestikerkennung

Erläuterung:

6−11: Elapsed time = Dauer, auch als ET abgekürzt zu finden.
11: Die ürsprüngliche Benennung von Smartkom wurde beibehalten und die Paradise−

Benennung als Ergänzung dazu gefügt.
3/4: Die Wörter ohne Symbole stellen die absolute Zahl der Vorkommen dar, während die vom Prozentzeichen gefolgte Wörter die Prozentzahl der Vorkommen im Turn oder im Dialog darstellen.
5: Die Nomenklatur wurde ersetzt.

Somit müssen Gestik– und „Mimik“eingaben sowie deren Erkennung bezüglich ihrer Dauer gemessen werden.
Dialogdauer, timeout, cooperativity und Barge In werden in ihrer Gesamtheit gewertet. Anstelle der in PARADISE vorgesehenen Vergleiche mit einem Standardsystem wird bei SmartKom die Stellung des Wichtels und die Synchronisierung von Stimme und Bewegungen evaluiert. Außerdem werden Gesichtspunkte wie N–Way Kommunikation (also gleichzeitig über verschiedene Modalitäten), Eingabekomplexität und Interaktionswege zur Bewertung der multimodalen Ein– und Ausgabemedien berücksichtigt. Dazu gehört auch der Bereich der Objekt Manipulationen bei der Gestikeingabe/Graphikausgabe.
Off–Talk, syntaktische bzw. semantische Korrektheit von Input/ Output, der semantische Gehalt der Ausgabe werden ebenfalls in die Auswertung mit einbezogen. Im sprachlichen Bereich wird außerdem noch auf die prosodischen Merkmale bei der Evaluation von Emotionsdaten eingegangen.

7.3 Lösungsansatz PROMISE

PROMISE löst diese Probleme durch die Verwendung von übergeordneten Konzepten (eindeutige Informationskluster s. 7.3.2.) abhängig von der Anwendung. Zur Berechnung der Task Success wird eine bipolare Funktion verwendet, die die möglichen Referenzantworten nicht mehr diffus werden läßt. Durch diesen Ansatz ist PROMISE unabhängig von statischen Referenzschlüsseln (s. 7.3.2).

7.3.1 Die Komplexität des Dialogs aufgrund von Multimodalität

- Wie können User–State–Erkennung und –Analyse bzw. Gestikerkennung und –analyse in Hinblick auf die Anforderungen der Benutzer optimal gewichtet werden?
- Durch die große Anzahl von multimodalen Dialogkomponenten, deren Funktionsweise oft untereinander ähnlich ist, können Interferenzen zwischen einzelnen Modalitäten auftreten (z.B. gleichzeitige Anwendung von Sprache und Gestik).
- Dabei muß bei der Evaluation berücksichtigt werden, inwieweit Eingaben tatsächlich gleichzeitig sind und ob diese tatsächlich auch gleichwertig in ihrer Intention sind und inwieweit das System diese multimodalen synchronen Eingaben tatsächlich verarbeiten kann.
- Durch die Gewichtung der Ergebnisse mit den entsprechenden Ergebnissen aus der Benutzerbefragung sowie durch die Einschätzung des semantischen Gehalts der Eingaben kann hierbei eine Bewertung erfolgen.
- Einschätzung der Benutzermeinung
  Zur Einschätzung der Benutzermeinung über das System werden verschiedene Fragetechniken verwendet. Neben den sog. offenen Fragen, die ausschließlich von Hand evaluiert werden, gibt es qualitative Bewertungen, mit einer 7wertigen Werteskala von "stimme komplett zu" bis "stimme gar nicht zu" reicht (s. 9.3). Die Antworten aus dieser Werteskala können div. Kosten aus der TE zugewiesen werden, um die Gesamtperformanz des Systems zu berechnen.

7.3.2Objektiv meßbare multimodale Qualitäts– und Quantitätsmaße

Nachfolgend findet sich eine Auflistung der in SmartKom verwendeten Qualitäts– und Quantitätsmaße, die objektiv gemessen werden können:

Qualitätsmaße	
system–cooperativity	measure of accepting misleading input
semantics	no.of multiple input
	possible misunderstandings
	of input/output
	semantical correctness
	of input/output
helps	no. of offered help for the actual interaction situation
recognition	speech
	facial expression
	gestures
transaction success	no. of completed sub–tasks
diagnostic error prompts	percentage of error messages
dialogue complexity	task complexity (needed information bits for one task)
	input complexity (used information bits)
	dialogue manager complexity (presentation of results)
ways of interaction	gestures/graphics vs.speech
	n–way communication (several modalities possible at the same time?)
synchrony	graphical and speech output
user/system turns	mixed initiative dialogue management
	incremental compatibility (sinngemäße Wiederholungen)

Quantitätsmaße	
barge–in	no. of user and system overlap by means of backchannelling, negation of output, further information
cancels	planned system interrupts due to barge–in
off–talk	no. of non–system directed user utterances

Quantitätsmaße	
elapsed time	duration of input of the facial expression
	duration of gestural input
	duration of speech input
	duration of ASR
	duration of gesture recognition
	mean system response time
	mean user response time
	task completion
	duration of the dialogue
rejections	error frequency of input which require a repetition by the user
timeout	error rate of output
	error rate of input
user/system turns	no. of turns
	no. of spoken words
	no. of produced gestures
	percentage of appropriate/inappropriate system directive diagnostic utterances
	percentage of explicit recovery answers

Die Generalisierung der einzelnen Kosten erfolgt ebenfalls über eine z–Normalisierungsfunktion:

$\mathrm{N}(x) = \frac{\bar{x} - x}{\sigma_x}$, mit σ_x Standardabweichung für *x und* $\bar{x}$ *Mittelwert.*

PROMISE gewichtet die Einzelkosten über eine Pearson Korrelation mit den jeweiligen Ergebnissen aus der Benutzerbefragung.

Die Performanz nach PROMISE ergibt sich zu:

$$\text{Performanz} = \alpha * \bar{\mathcal{T}} - \sum_{i=1}^{n} \omega_i * \mathrm{N}(c_i)$$

,wobei α ein Gewicht auf den Mittelwert von $\mathcal{T}$ (task success), die Kostenfunktionen c_i durch ω_i gewichtet werden und N ist eine Z–Wert Normalisierungsfunktion.
$\mathcal{T}$ kann die folgenden Werte erhalten:
$\mathcal{T}(j) = +1$: task success;
$\mathcal{T}(j) = -1$: task failure;

Im Gegensatz zu PARADISE kann in SmartKom eine erfolgreiche Task nicht über eine AVM über eine statische Menge von Referenzkeys berechnet werden. $\tau(j)$ erlaubt die Berechnung einer erfolgreichen Task durch die Verwendung eineindeutiger Informationskluster. Diese unterscheiden sich von Anwendung zu Anwendung und müssen von Hand definiert werden.

Z.B.:
Erfolgreicher Task Filmauswahl für den Abendfilm des heutigen Abends (Beginn 20h15):
"Stadt der Engel" 1 Informationskluster
20:15 Uhr VOX (= "Stadt der Engel") 2 Informationskluster
Spielfilm um 20:15 Uhr mit Meg Ryan und Nicolas Cage in den Hauptrollen
4 Informationskluster

Ein weiteres Problem von Paradise ist die stillschweigende Voraussetzung kooperativer User in dem Sinne, dass nur vollständige Tasks bewertet werden. Durch die Berechnung von $\tau(j)$ in PROMISE können auch unvollständige Tasks bewertet werden.

Für eine sinnvolle End-to-End-Evaluation müssen allerdings auch einige Einschränkungen gegenüber der "Freiheit" des Benutzers gemacht werden:

- Unterbindung von Störfaktoren wie Multiparty Benutzung
- komplette Abweichung von SmartKoms Anwendungsmöglichkeiten ("Stell' die Heizung an").

Würden derartige Störfaktoren bei der Berechnung der Performanz von SmartKom mit einbezogen werden, würde aufgrund von Faktoren wie "dialogue elapsed time" und "Antwortenmapping" der Dialog schlechter bewertet, obwohl die Reaktion des Systems angemessen war.

8 Was wird evaluiert?

8.1 Fragenkatalog zur technischen Evaluation

Bei der technischen Evaluation werden objektiv meßbare Kriterien aus den Aufnahmen, deren Annotationen und den dazugehörigen Logfiles extrahiert und statistisch bewertet. Zudem wird den Evaluierern ein Tool zur Verfügung gestellt, das alle End-to-End-Kriterien gegenüberstellt und somit eine Nachbewertung des Gesamtverhaltens erlaubt. Dabei werden Fragen wie:

- Welche Module funktionieren wie gut? (Trockentest & Schnittstelle Mensch-Module)
- Bearbeitungszeiten: Wie lange braucht die Verarbeitung einer Eingabe?
- Interaktionswege (mit welchem Modus antwortet das System auf verschiedene Eingabemodi des Anwenders)?
  - Spracheingabe -> Sprachausgabe
  - Gesteneingabe -> Sprachausgabe
  - Spracheingabe -> Displayausgabe
  - Gesteneingabe -> Displayausgabe
  - usw.

- Fehleranfälligkeit / Schwere der Fehler
- Frequenz der wiederholten Anfragen (wegen System, wegen User)
- Frequenz von erkanntem Offtalk

•Frequenz von Barge–In

•Bearbeitungszeiten von Eingabe bis Ausgabe

•Welche Hilfestellung bietet das System dem Anwender?
•Korrektheit der Ausgaben: Überwiegend falsche / korrekte Antworten (Oberfläche)

Alle objektiv meßbaren Evaluationskriterien genügen dem Paradise–Standard für Dialogsysteme. Je nach Szenario wurden diese auf Multimodalität angepaßt. Die nachfolgende Tabelle enthält prinzipiell zu evaluierenden Funktionen verschiedener Modi (aus Sprachverarbeitung (S), aus Gestikverarbeitung (G), aus Mimikverarbeitung (M)). Diese werden aber nicht bei allen Szenarien gleichstark vertreten sein!

Funktion	**Info aus files**	**Bewertung (global/turnweise)**	**Modus**
Antworten Rating	TRL/GES	global	Sprache (S), Gestik (G)
Anzahl der gesprochenen Woerter oder Gestikeingaben	TRL/ Video/Audio	turnweise	SG
Anzahl von Turns	TRL/GES/Log	Turnweise (eventweise)	SG
Aufnahmen	TE,EE	Global	
bargein (bargein%)	TRL/Log	Global(turnweise)	
Cancels (Cancels%)	Resets aus Logs	Global(Turnweise)	
DB	TE,EE	Global	
Dialogelapsed time (ET = Dauer)	TRL, LOG	global	
Dialogue Manager Complexity	Log	global	
elapsed time (ET = Dauer) der Gestikeingabe	GES/LOG	turnweise	G
elapsed time (ET = Dauer) der Gestikerkennung	GES/LOG	turnweise	G
elapsed time (ET = Dauer) der Spracherkennung	TRL/LOG	turnweise	S
elapsed time (ET = Dauer) der Sprecheingabe	TRL/LOG	turnweise	S
elapsed time (ET = Dauer) der „USMikeingabe“	USM/LOG	Turnweise	Mimik (M)

Funktion	Info aus files	Bewertung (global/turnweise)	Modus
elapsed time (ET = Dauer) der USMikerkennung	USM/LOG	Turnweise	M
Fehlerfrequenz von Anfragen (timeout (timeout%))	TRL/GES/USM	Global	S,G,M
Fehlerfrequenz von Antworten (rejections (rejections%))	TRL/GES/USM	Global	S,G,M
Fehlerrate von Anfragen	TRL/GES/(USM)	Global	S,G,M
Fehlertoleranz bei Anfragen (cooperativity)	TRL/GES/USM	Global	S, G, M
Gesture Recognition	GES	global	G
Helps (Help%)	Helpmodul/Log	Global(Turnweise)	
Incremental compatibility	TRL	global	Turnwiederholungen
Input Complexity	TRL/GES	global	S, G
Interaktionswege	Video/Audio/TRL	turnweise	S–G–Wechsel
Mean system response time	Log	turnweise	SG
Mean user response time	Log	turnweise	SG
Mixed initiative dialogue management	TRL/Video/Audio	global	SG
Number of senses	TRL/Video/Audio	global	S
N–Way Kommunikation	Video/Audio/Log	turnweise	SG–Wechsel
Objekt Manipulationen	Video/Audio/Log	turnweise	SG–Wechsel
Off–Talk	TRL/Log	turnweise	
Percentage of appropriate & inappropriate system direcctive & diagnostic utterances	TRL	Turnweise	
Percentage of diagnostic error messages	Log	Global	

Mixed initiative dialogue management	TRL/Video/Audio	global	SG
Percentage of explicit recovery answers	TRL	Turnweise	
Percentage of implicit recovery answers	TRL	Turnweise	
Performance	Log, TRL/ GES/USM	global	Minimaler Aufwand
possibility of misunderstanding of input/output	TRL/Video/Audio	global	S
Prosodische Merkmale	TRL	turnweise	S
Reliability	Log	global	Software, Hardware
Schwierigkeit des Dialogs	TRL/Fragebogen	Global	SG
Semantical Correctness of Input/ Output	TRL	turnweise	S
Semantischer Gehalt der Ausgabe	TRL	global	
Speech Recognition	TRL	global	S
Synchronisierung	Video/Audio–Synthese/Log	turnweise	SGM
Syntactical Correctness of Input/ Output	TRL	turnweise	S
System gefuehrte Dialoge	TRL/Video/Audio	global	SG
Task Completion	Video/Audio/TRL/LOG	global	SGM
Task Complexity	TRL/GES	global	SG
Transaction Success	TRL/GES/Log	turnweise	SGM
User gefuehrte Dialoge	TRL/Video/Audio	global	SG
USMimikerkennung	USM	global	M
Zeitverschiebungen	Video/Audio/mars/ Log	turnweise	Segmentgrenzen

Die Bewertung erfolgt global oder turnweise. Die Herkunft der entsprechenden

Information ist gem. Filekürzel erklärend.

8.2Fragenkatalog zur ergonomischen Evaluation

Die ergonomische Evaluation erfolgt zweiteilig; einige Variablen werden vor Beginn des Experimentes erhoben, um dann mit anschließend erhobenen Variablen kontrastiert werden zu können. Vor Beginn des Experimentes wird z.B. die Einstellung zur Emotionsanalyse abgefragt. Dies wird dann mit einer post−experimentellen Frage zum selben Thema kontrastiert um mögliche Einstellungsänderungen durch die Nutzung von SmartKom eruieren zu können. Die Nummern in Klammern nach der Beschreibung beziehen sich auf den Fragebogen, Version vom 04.03.02, siehe auch Anhang B.

Zu den vor dem Experiment erhobenen Variablen gehören:

Teil I **„Allgemeine Fragen zur Person“**

1.)Allgemein die Zugangsmöglichkeit zu einem Computer (1).
2.)Die Nutzungshäufigkeit eines Computers (2).
3.)Die selbsteingeschätzte Technik−Kompetenz (3).
4.)Abfrage Voreinstellung über Emotions−Analyse und Aufnahme durch eine Kamera (4).
5.) Präferierte Ansprache durch das System (5)

Die folgenden Fragen werden NACH erfolgtem Test des SmartKom−Systems gestellt:

Teil II **„Bedienung des Systems SmartKom“**

1.)Der Paradise−Fragekomplex und einige zusätzliche Features von SmartKom (6).
2.)Abfage einzelner Aspekte des SmartKom−Systems nach Funktionalität (7).
3.)Frage nach Schwierigkeiten bei Bedienung von SmartKom (8).
4.)Frage nach Verbesserungsmöglichkeiten von SmartKom mit Begründung (9).
5.)Waren Kommunikationsmöglichkeiten mit SmartKom ausreichend (10).
6.)Welche zusätzlichen Geräte werden gewünscht (11).
7.) Einschätzung der Kommunikation mit SmartKom als menschlich oder eher maschinell mit Begründung (12)
8.) Abfrage nach Änderung der Einstellungen zur Emotions−Analyse und Aufnahme durch die Kamera (13).

9.) Schwierigkeit in der Rolle zu bleiben mit Begründung (14)

10.) Abfrage der präferierten Identifikationstechnik (15)

Teil III **„Mögliche Benutzergruppen“**

1.)Einschätzung der primären Benutzergruppen (16).
2.)Abfrage des „Gefallens“ einzelner Aspekte von SmartKom (Displayein−/ausgabe, Sprachein−/ausgabe, Möglichkeit der menschlichen Interaktion, Design insgesamt, Reaktion auf Emotionen) (17).
3.)Allgemeine Frage nach negativen Aspekten von SmartKom (18).
4.)Abfrage nach potentiellem Wert des Systems (19).
5.)Abfrage von schlecht durchgeführten Aktionen (20).
6.)Abfrage von gut durchgeführten Aktionen (21).
7.)Abfrage von Gefühlen während des Experiments mit Begründung (22).
8.)Bewertung einzelner Elemente von SmartKom nach deren Gehalt an Innovation, Qualität, High−Tech, Fun, Nützlichkeit (Displayein−/ausgabe, Sprachein−/ausgabe, Möglichkeit menschlicher Interaktion, Design insgesamt) (23).
9.)Frage nach Elementen, die benutzerspezifisch angepaßt werden können sollten (24).
10.) Bitte SmartKom mit einer Automarke zu vergleichen (25).

Teil IV **„Bewertung der Persona Smartakus“**

1.) Bewertung der Kompetenz von Smartakus mit Begründung (26).
2.) Bewertung der Natürlichkeit der Aktionen der Persona (27).
3.) Bewertung des Aussehens der Persona (28).
4.) Abfrage von positiven Aspekten der Persona (29).
5.) Abfrage von negativen Aspekten der Persona (30).

Einige Fragen in eigener Sache
1.) Frage nach Schwierigkeiten mit dem Fragebogen mit Begründung (31)
2.) Möglichkeit, dem Team etwas mitzuteilen (32)

Um die einzelnen Szenarien−Unterschiede gezielt abfragen zu können, werden bei einzelnen Fragen Szenarien−abhängige Aspekte mit eingebaut. Diese hängen dem Fragebogen−Dokument am Ende an und werden jeweils zu Frage 6, 17, 23 und zusätzlich als Frage 33 integriert. Die szenario−spezifischen Fragen werden je nach Bedarf ergänzt oder verändert. <siehe Beispiel oben und Frage 4 und 17.>

8.3 Übersicht über das Evaluationstool

8.3.1 Anforderungen

Der primäre Einsatz des Evaluationstools liegt in einer quasi objektiven Beurteilung technischer wie ergonomischer Evaluationsergebnisse durch Evaluierer unter Heranziehung aller zur Verfügung stehenden Video−, Audio− und Annotationsdaten. Das Tool bietet eine Möglichkeit, entsprechende Videodatenströme gezielt durch Anklicken der entsprechenden Zeiträume vorwiegend unter Linux abzuspielen. Durch die vorgegebenen Entscheidungsmöglichkeiten (Gewicht auf TE, Gewicht auf EE, ausgewogen) können Endergebnisse auch bei einer großen Anzahl von Evaluierern gut miteinander verglichen werden, da sowohl die Meinung der Evaluierer als auch die der VPn standardisiert wird. Durch die Bereitstellung eines Kommentarfelds haben die Evaluierer die Möglichkeit, nicht abgedeckte bzw. eingeschränkte Auffälligkeiten zu kommentieren.

Die nachfolgende Tabelle gibt einen Überblick über die zu vergleichenden Werte aus der TE und EE.

Qualitäts− und Quantiätsmaße	Ergonomische Frage
Transaction success Task complexity	Die Aufgabe war einfach zu lösen
Misunderstanding of input Offtalk Misunderstanding of output	SmartKom hat mich verstanden
Semantical, syntactical correctness Incremental compatibility	SmartKom ist einfach zu verstehen

Qualitäts– und Quantiätsmaße	Ergonomische Frage
Mean system response time	SmartKom hat zum größten Teil passende Antworten gegeben Die Geschwindigkeit des Systems war angemessen
Mean user response time Timeout	Ich wußte während des Dialogs immer, was ich sagen konnte
Acc. gesture recognition	Die Eingabe über Zeigen hat funktioniert
Acc. ASR	Die Eingabe über Sprache hat funktioniert
Dialogue complexity	SmartKom hat so funktioniert wie ich es erwartet hatte SmartKom reagierte schnell auf meine Anfragen Smartkom ist unkompliziert zu bedienen bot in Anzahl und Qualität angemessene Informationen an
Percentage of appropriate/inappropriate system directive diagnostic utterances Percentage of explicit recovery answers	Smartkom ist unkompliziert zu bedienen
repetitions No. of ambiguities Diagnostic error messages Rejections Timeout	benötigt Eingaben nur einmal, um zum Ziel zu gelangen
Help–analyzer	bietet eine gute Hilfefunktion an
Output complexity (display)	hat ein übersichtlich gestaltetes Display
Mean elapsed time Task completion time Dialogue elapsed time	reagierte schnell auf meine Aufforderungen (allgemein)
Duration of speech input Duration of ASR	reagierte schnell auf gesprochene Befehle
Duration of gestural input Duration of gesture recognition	reagierte schnell auf Zeigen
BargeIn Cancel	läßt mich unterbrechen
Dialogue complexity	War die Aufgabe schwierig oder hatten Sie Schwierigkeiten in der Rolle zu bleiben?

Qualitäts- und Quantiätsmaße	Ergonomische Frage
Gesture turns Ways of interaction	Displayeingabe (allgemein)
Display turns	Displayausgabe (allgemein)
Speech input	Spracheingabe (allgemein)
Speech synthesis (synchronicity)	Sprachausgabe (allgemein)
N-way communication Ways of interaction Error rate of questions Input complexity	Die Möglichkeit mit dem System „menschlich"zu interagieren
Recognition/duration of facial expression Prosodic features	SmartKom reagiert auf meine Emotio-nen
Synchronicity Graphical output (turns) Cooperativity	Wie beurteilen Sie die Kompetenz von Smartakus? Wie natürlich wirkten die Aktionen der Persona auf Sie?
Gestural input	Gestikeingabe

Neben der eigentlichen Funktion der Evaluation bietete dieses Tool auch die Möglichkeit, Userstate und Gestik zu annotieren. Ebenfalls können durch seine Verwendung auch die verschiedenen Erkennerergebnisse gewichtet werden. Außerdem bietet das graphische Interface auch eine gute Möglichkeit, die Evaluationsergebnisse für die Entwickler anschaulich zu präsentieren. Durch die Wertung der Evaluierer ist auf einen Blick ersichtlich, auf welchen Aspekt der Evaluierung bei der Verbesserung der Module im Einzelfall mehr Gewicht liegen sollte.

8.3.2 Realisierung

Die Realisierung erfolgte in zwei Schritten:
Zuerst mußte eine Möglichkeit gefunden werden, Teile von Videoströmen unter Linux abzuspielen. Dazu wurde der Videoplayer xanim dementsprechend umprogrammiert. Die Ansteuerung erfolgt dann direkt aus dem Java-Evaluationstool, dessen Funktionalität bereits oben beschrieben wurde.

8.3.2.1 Anbindung xanim

XAnim wird durch zwei Wrapper angebunden: embedded_xanim und XAnimRemoteControl.
Die Ansteuerung von Teilen eines Videos erfolgt über vier neu einprogrammierte Befehlsoptionen. Dabei besteht die Möglichkeit, Zeiten (in frames oder Milisekunden) auch per Hand einzugeben.

- embedded_xanim ist ein interaktives Interface für XAnim programmiert in C. Es gibt ein Fenster in spezifizierter Größe und Position aus. XAnim wird über Kommandozeilenargumente angesteuert, um das gewünschte Video abzuspielen. Das

Video kann an jeder Stelle durch den "Stop"–Befehl unterbrochen werden, ohne daß es zum Anfangspunkt (wie in der unveränderten XAnim– Version) zurückspringt.

- XAnimRemoteControl ist eine höherwertige Java Wrapperklasse für embedded_xanim. Bei der Initialisierung wird eine Instanz von embedded_xanim gestartet. Videofiles können über die Play–Methode abgespielt, wiederholtes Abspielen kann durch die Stop–Methode beendet werden. Die Quit–Methode verläßt das Programm.

8.3.3 Datenbank

Alle Daten, die vom Evaluationstool verwaltet werden, kommen aus einer MySQL Datenbank. Es handelt sich hierbei um die Daten der TE (aus Systemlogfiles und Annotationen) und die Daten aus der EE (Fragen und Antworten aus dem Fragebogen). Beide Datenbereiche werden über eine Tabelle verlinkt. Die Ergebnisse des Evaluationsschritts werden ebenfalls in die Datenbank – in eine eigene, jedoch verlinkte Synthesetabelle – zurückgeschrieben.

8.3.4 Oberfläche

Die Evaluierung der Daten erfolgt über eine plattformunabhängige in Java geschriebene GUI. Die graphische Oberfläche ist primär unterteilt in drei Teilbereiche: ein Java Evaluationsfenster, ein XAnim–Fenster für das Abspielen von Videos und ein xemacs–Fenster für die Annotationen.
Das Javafenster ist wiederum unterteilt in drei gleich große Felder:

- Das linke Feld stellt die jeweiligen Ergebnisse der TE gem. oben beschriebener Zuordnung dar. Diese Daten werden aus der MySQL Datenbank pro Session extrahiert.
- Das rechte Feld enthält die zugeordneten EE–Ergebnisse aus der Datenbank exportiert.

- Im mittleren Feld befinden sich die vordefinierten Synthesekategorien sowie das Kommentarfeld. Diese Daten werden nach Bearbeitung in die MySQL Datenbank exportiert.

1. Beim Start der GUI mit einer Session wird gleichzeitig der alignierte XAnim mit dem entsprechenden Sessionvideo und ein aligniertes xemacs–Fenster mit den jeweiligen Annotationen aufgerufen.

Während der Evaluation können die Evaluierer durch das Videofile navigieren durch anklicken bestimmter Zeitmarken im TE–Feld bzw. durch freie Eingaben an jeder Stelle der Evaluation.

9 Literatur

[SKantrag] DFKI et al.: Leitprojekt SmartKom – Dialogische Mensch–Technik–Interaktion durch koordinierte Analyse und Generierung multipler Modalitäten

[E2EVM00] S.Jekat, L. Tessiore: End–to–End Evaluation of Machine Interpretation Systems: A Graphical Evaluation Tool. In: Proceedings of LREC 2000, Athen Mai 2000.

[1] M.A. Walker, et al.: DARPA Communicator Dialog Travel Planning Systems: The JKune 2000 Data Collection, LREC 2000.

[2] M. Rahim et al.: VOICE–IF: a mixed–initiative spoken dialogue system for AT&T conference services. EUROSPEECH 2001

[3] M. Rogati et al.: Training a Sentence Planner for Spoken Dialog: The Impact of Syntactic and Planning Features. EUROSPEECH 2001

[4] M.A. Walker et al.: Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies,In Computer Speech and Language, Vol. 12, No. 3, 1998.

[5] M.A. Walker, D.J. Litman, C.A. Kamm, A. Abella: PARADISE: A Framework for Evaluation Spoken Dialogue Agents. Annnual Meeting of the Association of Computational Linguistics. ACL 1997.

[6] U. Andenfilger: Some Remarks about the Validation of Information Systems Development. In Interdisciplinary Foundation of Systems Design and Evaluation. Seminar Report 97 of International Conference and Research Center for Computer Science, edited by R. Keil–Slawik, Saarbrücken Schloss Dagstuhl, University of Saarland.

[7] L. Hirschmann, D.A. Dahl, D.P. McKay, L.M. Norton, M.C. Linebarger: Beyond class A: A proposal for automatic evaluation of discourse. In Proceedings of the Speech and Natural Language Workshop,1990.

[stat. Bundesamt]

[8] V. Penide: ATTRIBUTE VALUE MATRIX. Detaillierte Beschreibung zur Berechnung der AVMs. ~smart/EVAL/AVM/AVM–paradise.sdw

10 Anhang A: Zeitplan Eval

Zeitplanung Evaluation – Version 2.1

Nachdem auf der Moko (13.11.) essentielle Anpassungswünsche bzgl. unseres Zeitplans geäußert wurden und um trotzdem möglichst viel von unseren Vornahmen vom Eval–Workshop zu verwirklichen, stellen wir Euch hier die revidierte Form unseres Zeit– und Ressourcenplans für die Evaluation vor.

DIESER VORSCHLAG WIRD – SOFERN KEIN FEEDBACK VON EUCH KOMMT – IN KURZFASSUNG DEM LENKUNGSKREIS ZUR VERBINDLICHEN ENTSCHEIDUNG VORGELEGT!

Die Änderungen gegenüber der letzten Fassung betrifft vorwiegend die Evaluationszeiträume.

Anstelle der bis jetzt geplanten aussschließlich abschließenden Evaluation tritt eine Mischform aus entwicklungsbegleitender Evaluation und abschließender Evaluation, derart, daß wir bis zur CeBit 2003 (die von euch angestrebte Deadline des Prototypen) die drei Szenarien in kürzeren Zeiträumen evaluieren. Dazu verwenden wir den zum Beginn unserer Evaluationsphase aktuellsten Demonstrator.

Danach gibt es pro Szenario einen kürzeren Evaluationszeitraum mit dem tatsächlichen Endsystem (invariabel für die Eval ab CeBit). Das hat gegenüber der ersten Variante folgende Vorteile:

- kürzere Zeiträume
- 2 SK–Versionen
- laufender Feedback in die Module
- wiss. Endergebnisse
- sichtbare Fortschritte
- Endsystem
- bessere Akzeptanz durch Partner

Wir haben uns damit noch mehr vorgenommen und brauchen, damit die Eval sowohl für uns, als auch für euch optimal über die Bühne geht, Eure Hilfe mehr denn je! Unter benötigte Ressourcen steht, bei was genau wir auf Euch angewiesen sind. Ganz wichtig:

Vom DFKI wünschen wir uns jemanden, der jeweils in der Vorbereitungsphase für einige Zeit vorbeischaut und uns hilft, das Testbed zum Laufen zu bekommen.

Beim Zeitpunkt “Beginn der Vorbereitung”muß jeweils ein weitgehend funktionstüchtiges System vorhanden sein!

Und zusätzlich brauchen wir eine email–hotline, bei der wir mindestens 2x täglich Lösungen unserer Probleme von der Systemgruppe erhalten!

Am Ende des Dokuments findet Ihr jeweils noch ein paar Worte zu den einzelnen Punkten. Alle Änderungen gegenüber der ersten Version sind markiert. Die wichtigsten Vorschläge vorab:

Standardprogramm ist in jedem Szenario die Befragung laut Antrag sowie eine technische Evaluation der **Audio–Daten.**

In **Public** bezieht die techn. Evaluation zusätzlich die Video–**Daten** mit ein, bei **Mobil** gibt es die gewünschte “**Langzeit–Untersuchung**” und in **Home** einen zusätzlichen **qualitativen Teil**.

Die Evaluation beruht soweit möglich auf den Paradise Kriterien, die erweitert wurden, um den Besonderheiten von SmartKom gerecht zu werden (**erweiterte Paradise Evaluation SmartKom**).

Je Evaluation werden die szenariobedingten Probleme (vorwiegend inhaltliche), Teilergebnisse, Logfiles etc. direkt zur Verfügung gestellt. Der Zugriff erfolgt für die jeweiligen Szenarioverantwortlichen und die Systemgruppe DIREKT über Paßwort.

Daneben werden – wie geplant – allgemeine Probleme mit Modulen, Testbed etc. direkt über die Moko–Liste gepostet.

Bitte laßt Euch unseren Vorschlag durch den Kopf gehen und sagt uns Bescheid, was Ihr gut findet und was Ihr gerne wie ändern möchtet. **Auf der nächsten Projektlenkungssitzung erwarten wir einen verbindlichen Beschluß von Euch!**

Entwicklungsbegleitende Evaluation:

	Home	**Public**	**Mobil**
Beginn Vorbereitung	Okt 2001	Apr 2002	Juli 2002
Eval–Zeitraum	April 02 – Juni 02	Juli 02 – September 02	Okt 02 – Jan. 03 (wg. Weihnachtsloch)
Standardpro–gramm	– 30 VP/9 Min – **Audio** Paradise+ – Befragung/ Interview	– 30 VP/9 Min – **Audio** Paradise+ – Befragung/ Interview	– 30 VP/9 Min – **Audio** Paradise+ – Befragung/ Interview
Zusätzlich	– 5 Vpn mit ausführlichem offenem Interview	– T–Eval: **Video** Paradise+	– 5 Vpn mit 5 Sessions
benötigte Ressourcen	* Hilfe Einrichtung Testbed (DFKI) * Fernseher (Philips)° * Videorecorder (Sony)° * Webpad (Philips) * Simulation Sprecherverif. * Funkmikro ° Simulation	* Hilfe Einrichtung Testbed (DFKI) * Simulation Objektkamera (Siemens) * Simulation Unterschrift (Siemens?)	* Hilfe Einrichtung Testbed (DFKI) * Ipaq (Daimler) * Funkmikro (N.N.)

Abschließende Evaluation:

	Home	**Public**	**Mobil**
Beginn Vorbereitung	Jan 2003	März 2003	Mai 2003
Eval–Zeitraum	März 03 – April 03	Mai 03 – Juni 03	Juli 03 – August 03

	Home	Public	Mobil
Standardpro–gramm	– 20 VP/9 Min – **Audio** Paradise+ – Befragung/ Interview	– 20 VP/9 Min – **Audio** Paradise+ – Befragung/ Interview	– 20 VP/9 Min – **Audio** Paradise+ – Befragung/ Interview
Zusätzlich	– 5 Vpn mit ausführlichem offenem Interview	– T–Eval: **Video** Paradise+	
benötigte Ressourcen	* Hilfe Einrichtung Testbed (DFKI) * Fernseher (Philips) * Videorecorder (Sony) * Webpad (Philips) * Simulation Sprecherverif. * Funkmikro	* Hilfe Einrichtung Testbed (DFKI) * Simulation Objektkamera (Siemens) * Simulation Unterschrift (Siemens?)	* Hilfe Einrichtung Testbed (DFKI) * Ipaq (Daimler) * Funkmikro (N.N.)

Benötigte Ressourcen

Hilfe Einrichtung Testbed für jeden Durchgang: Siehe Abschnitt Zeitplanung

Bezüglich der Geräte hatten wir ja schonmal grob ausgemacht, wer uns was leiht. Auch die Geräte brauchen wir jeweils 3 Monate vorher, um sicherzustellen, daß alles bis zum Beginn des Eval–Zeitraums klappt. Da wir mit Home anfangen wollen, brauchen wir die Geräte entsprechend so bald wie möglich!

Für Home werden in der entwicklungsbegleitenden Evaluation die Ansteuerung von Fernseher und VCR simuliert. Die entsprechenden Geräte werden von der Datensammlung übernommen. Trotz der geänderten Zeiträume wird das Webpad dennoch über die gesamten 5 Monate benötigt. Der Geräteverleih erfolgt über einen Zeitraum von ca. 5 Wochen vom DFKI, Sony und Philips in folgender Reihenfolge:

Philips(1.2.02 – 8.03.02)

DFKI(11.3.02 – 16.04.02)

Sony(17.4.02 – 23.05.02)

DFKI (24.5.02 – 30.06.02)

Eine Kürzung der Dauer kann nur insofern erfolgen, wenn

das Webpad funktionsbereit ist (d.h. alle erforderlichen Installationen werden von den jeweiligen Partnern bereits durchgeführt)

das Webpad durch den jeweiligen Partner hier installiert wird

Der vorgeschlagene Zeitraum von 3 Wochen ist für die LMU inakzeptabel, da zu viel Zeit von wertvoller Aufnahmezeit verloren geht. Eine Alternative wäre, wenn die entsprechenden Partner die Geräte dann untereinander austauschen.

Abschließende Evaluation Home: Hier werden ALLE Geräte für den Evaluationszeitraum von 2 Monaten benötigt. Eine Ausleihreihenfolge wird zu gegebener Zeit aktualisert!

Simulation Objektkamera: Auf dem Eval–WS haben wir vereinbart, daß die Objektkamera das Fax nicht wirklich aufnimmt (weil die Vp dies eh nicht merkt), sondern daß das Module mit einem „Simulationsmodus“ ausgestattet wird. Zum Testen brauchen wir diesen Modus wie alles andere zum Vorbereitungsbeginn – sprich im April 02.

Selbiges gilt für die Biometrie (Unterschrift und Sprecherverifikation) – wir wollen auch hier ein 100%iges Funktionieren simulieren (für die Befragung). Der Simulationsmodus für die

Sprecherverifikation muß bis Okt 01 da sein, der für die Unterschrift bis April 02.

Woher wir das Funkmikro und ev. Geräte/Ressourcen bekommen, die wir vergessen haben, ist noch offen.

WICHTIG:
Die Evaluation setzt IMMER ein funktionierendes System voraus! Dafür müssen allerdings einige Funktionen (Biometrie wg. Training – Naivität) bzw. Hardware (Objektkamera wg. div. Grabberkarten; TV/Videorecorder wg. Verleihzeiträumen) simuliert werden!

Eine andere Lösung gäbe es nur dann, wenn von den <u>jeweiligen Partnern sowohl das entsprechende Gerät bzw. die entsprechende Funktion zur Verfügung gestellt wird, als auch eine Vor–Ort–Installation durch den entsprechenden Partner (entsprechend rechtzeitig VOR der Eval) inklusive einer Fehlerhotline (für den gesamten Evaluationszeitraum) erfolgt</u>. Solange diese Bedingung nicht gewährleistet werden kann, muß simuliert werden oder die entsprechende Funktion /das entsprechende Gerät wird gar nicht evaluiert.

Standardprogramm

In allen Szenarien wird wie im Antrag vorgesehen eine Befragung durchgeführt, jeweils mit einer möglichst breit gestreuten Benutzergruppe (s.u.). Zusätzlich machen wir in allen Szenarien eine technische Evaluation (des jeweiligen Gesamtsystems), die an der Paradise Norm orientiert ist, soweit sich die Kriterien auf ein multimodales System übertragen lassen, bzw. fügen einige neue Kriterien (die speziell auf SmartKom bezogen sind) hinzu. Für diese technische Evaluation werden die Audiodaten komplett aufgenommen und transliteriert.

Relevante Ergebnisse (v.a. der abschließenden Evaluation) sowie Evaluationsmethoden werden veröffentlicht.

Alle Daten werden aufgenommen, ggf. jedoch nicht annotiert bzw. evaluiert *(z.B. gibt es bei der AudioParadise+ zwar die Mimik– und Gestikaufnahmen, sie werden jedoch nicht annotiert und bei der Evaluation nicht berücksichtigt.)*

Stichprobe

Die Probanden werden in Bezug auf **Alter** und **Geschlecht** nach dem Bevölkerungsdurchschnitt (laut stat. Bundesamt) rekrutiert. Zusätzlich werden als „weiche" Kriterien die Nationalität und der Schulabschluß berücksichtigt, in dem jeweils ein Mindest– bzw. ein Maximalwert eingehalten werden.

10% 14–21 Jahre

32% 21–40 Jahre

32% 40–60 Jahre

26% 60+ Jahre

52% Frauen

48% Männer

mind. 4%, max. 10% **Personen ohne deutschen Paß** (10% ist der Bevölkerungsdurchschnitt)

„Personen ohne deutschen Paß" $\neq$ Akzentvariation der VPn

Der Anteil der VPn mit Akzent wird in etwa dem der WOZ–Aufnahmen entsprechen.

Schulabschluß:

Volksschule/Hauptschule:	mind. 25%, max. 50%	(50% BV–Schnitt)
Realschule:	mind. 28%, max. 33%	(28% BV–Schnitt)
(Fach)Hochschulreife:	mind. 22%, max. 33%	(22% BV–Schnitt)

Befragung/Interview:
Wie im Antrag vorgesehen wird in jedem Szenario eine Befragung durchgeführt. Der überwiegende Teil der Befragten wird einen standardisierten Online−Fragebogen beantworten, der wiederum aus einem allgemeinen Teil besteht, der in allen Szenarien gleich ist (und entsprechend Vergleiche und Zusammenfassungen zwischen den Szenarien ermöglicht) und aus einem szenario−spezifischen Teil. Einige wenige Probanden werden für das Szenario Home mithilfe eines umfassenden, offenen Interview befragt, welches Aspekte abfragt, die mit einem Fragebogen nicht erfaßt werden können: (emotionale) Einstellung des Benutzers, Wirkung des Systems, kritische Aspekte wie Hemmschwellen, Sicherheitsbedürfnisse.

– Online−Fragebogen mit allgemeinem Teil (identisch bei allen Szenarien) und Szenario−spezifischem Teil

– Interview mit einigen offenen Fragen zur qualitativen Auswertung

−> subjektive Effizienz und Akzeptanz (statist. Daten, qualitative Reports)

Audio Paradise+

Zusätzlich zu der im Antrag vorgesehenen Befragung wird in jedem Szenario eine technische Evaluation durchgeführt. Dazu werden die Daten komplett aufgenommen (Audio) und transliteriert. Abgeprüft werden die Paradise−Kriterien, die sich auf ein multimodales System übertragen lassen und einige zusätzliche, SmartKom−spezifische Kriterien. Ergebnisse sind die hard facts der Funktionstüchtigkeit und objektiven Effizienz des Systems.

– Audioaufnahme und Transliteration komplett

– Paradise−Kriterien

– SmartKom Kriterien

−> Benutzbarkeit/Funktionstüchtigkeit und objektive Effizienz (Fehlerzahlen, Prozentwerte etc.)

Zusatzprogramm

Um auf besondere Wünsche einzugehen und den SmartKom spezifischen Besonderheiten gerecht zu werden, fügen wir bei jedem Szenario noch ein Extra hinzu:

Bei Public eine auf Mimik und Gestik ausgeweitete techn. Evaluation, weil diese Eingabemodi hier eine besondere Rolle spielen.

Bei Home eine extra Runde Interviews, weil Ihr sie Euch gewünscht habt und weil sie inhaltlich sehr interessante Daten liefern.

Bei Mobil die gewünschte „episodische" Untersuchung, wie und ob sich die Einstellung der Benutzer ändert, wenn sie SmartKom häufiger benutzen.

Auch hier werden die Ergebnisse veröffentlicht.

Public: T−Eval: Video Paradise+

Zusätzlich zur normalen technischen Evaluation werden im Public Szenario auch die Video Daten komplett aufgenommen und gelabelt, weil hier die Mimik und Gestenerkennung eine besondere Rolle spielt. Abgeprüft werden wie bei den Audio−Daten die an SmartKom angepaßten Paradise−Kriterien. Die Abweichungen von Paradise sind hier etwas größer, weil Paradise auf *Sprach*−Dialogsysteme ausgerichtet ist. Ergebnisse sind wie bei der tech. Eval Audio die hard facts der Funktionstüchtigkeit und objektiven Effizienz des Systems.

– Mimikaufnahme und User−State Labelung komplett

– Sivit−, Beamer−, Seitenaufnahme und Gestiklabelung komplett

– weitere Paradise−Kriterien

– weitere SmartKom Kriterien

Home: 5 Vpn mit ausführlichem offenem Interview

Mit einigen Vpn wird eine ausführliche Tiefenexploration durchgeführt, mit dem Ziel, Fragen zu beantworten, die sich mit einem Fragebogen und Durchschnittswerten nur schwer erfassen lassen: Wie wirkt das System (qualitativ hochwertig, innovativ, freundlich, interessant etc. oder eben nicht)? Welche Einstellung hat der Benutzer (eher positiv oder eher negativ und warum)? Welchen emotionalen Bezug hat er zu dem System, wie fühlt er sich bei der Benutzung (nervös, geborgen, interessiert, genervt – und warum)? Welche Benutzungs–Hemmschwellen gibt es? Würde er es tatsächlich benutzen? Wie kommt er mit kritischen Aspekten zurecht, z.B. der Kamera, Unterschrifterkennung, Aspekten der Sicherheit?

Wie kommen die VPn mit lean–back/lean–forward–Modi zurecht? Wo gibt es Probleme? Wie stellen sich die VPn solche Transitionen idealerweise vor?

Die ursprünglich 10 Vpn wurden zu gleichen Teilen auf entwicklungsbegleitende und abschließende Evaluation aufgeteilt. Um der 2/3–1/3– Aufteilung allerdings weitgehend treu zu bleiben, gibt es nachfolgedes Zusatzprogramm nur bei der entwicklungsbegleitenden Evaluation.

Mobil: 5 Vpn mit 5 Sessions

Um abzuschätzen, ob sich bei wiederholter Benutzung des Systems Einstellungsänderungen ergeben werden ausgesuchte Vpn öfter aufgenommen und mit einem speziellen Fragebogen/Interview befragt.

Mobil: Visualisierung der Userreaktion

11 Anhang B: Fragebogen zur Ergonomischen Evaluationen

Fragebogen zum System SmartKom

Teil I: Bitte beantworten Sie zunächst einige allgemeine Fragen zu Ihrer Person:

1.) Haben Sie Zugang zu einem Computer? OJ ON
(bei N weiter mit Frage 3)

2.) Wie häufig benutzen Sie einen Computer?

mehrmals täglich	
täglich	
mehrmals wöchentlich	
einmal wöchentlich	
mehrmals im Monat	
einmal im Monat	
weniger als einmal im Monat	
selten	
gar nicht	

3.) Wie schätzen Sie Ihre Kompetenz auf folgenden Gebieten ein:

	hoch					**niedrig**	
	+++	++	+	+/–	–	––	–––
Software Programmierung							
Computerbedienung							
Nutzung des Internets							
Bedienung der Stereoanlage							
Bedienung elektrischer Küchengeräte							
Bedienung eines Systems über Spracheingabe							

4.) Ihnen wurde mitgeteilt, daß SmartKom Ihre Emotionen mithilfe einer Kamera aufnimmt und sie analysiert. Finden Sie diesen Aspekt eher hilfreich, oder unangenehm? Warum?

____

____

____

____

__

5.) Wie würden Sie gerne vom System angesprochen werden? Mit Du, Sie, oder ist es Ihnen egal?

O Du O Sie Oegal

Bitte wenden Sie sich nun dem System SmartKom zu und beantworten die nachstehenden Fragen erst NACH erfolgter Bedienung.

Teil II: Bitte beantworten Sie zuerst diese allgemeinen Fragen zum Thema „Bedienung des Systems SmartKom“.

6.) Bitte geben Sie an, wie sehr Sie folgenden Aussagen zustimmen:

	stimme zu					**stimme nicht zu**	
	+++	++	+	+/–	–	––	–––
Die Aufgabe war einfach zu lösen							
SmartKom hat mich verstanden							
SmartKom ist einfach zu verstehen							
SmartKom hat zum größten Teil passende Antworten gegeben							
Die Geschwindigkeit des Systems war angemessen							
Ich wußte während des Dialogs immer, was ich sagen konnte							
Die Benutzung von SmartKom wäre für mich eine Erleichterung							
Die Eingabe über Zeigen hat funktioniert							
Die Eingabe über Sprache hat funktioniert							
Ich würde SmartKom gerne nutzen							
SmartKom hat so funktioniert wie ich es erwartet hatte							
SmartKom reagierte schnell auf meine Anfragen							

7.) Unterstützt SmartKom die Erledigung Ihrer Aufgabe, ohne Sie unnötig zu belasten?

SmartKom	+++	++	+	+/–	–	––	–––	
ist unkompliziert zu bedienen								ist kompliziert zu bedienen
bot in Anzahl und Qualität angemessene Informationen an								bot in Anzahl und Qualität unangemessene Informationen an
benötigt Eingaben nur einmal, um zum Ziel zu gelangen								benötigt viele Eingaben mehrmals, um zum Ziel zu gelangen
bietet eine gute Hilfefunktion an								bietet keine ausreichende Hilfefunktion an
hat ein übersichtlich gestaltetes Display								hat ein unübersichtlich gestaltetes Display
reagierte schnell auf gesprochene Befehle								reagierte zu langsam auf gesprochene Befehle
reagierte schnell auf Zeigen								reagierte zu langsam auf Zeigen
läßt mich unterbrechen								läßt mich nicht unterbrechen

8.) Welche Schwierigkeiten gab es bei der Bedienung von SmartKom?

9.) Welche Elemente von SmartKom sollten verbessert werden? Warum?

10.) Bei SmartKom haben Sie die Möglichkeit, Ihre Befehle über Spracheingabe oder direkt am Display zu übermitteln. Waren die Ihnen zur Verfügung gestellten Möglichkeiten mit SmartKom zu kommunizieren ausreichend?
O ja O nein O weiß nicht

11.) Wenn nein, welche zusätzlichen Geräte wünschen Sie sich:
- O Telefonhörer
- O Tastatur
- O Maus
- O Drucker
- O sonstiges ____

12.) Hatten Sie das Gefühl, eher mit einem Menschen oder eher mit einer Maschine zu kommunizieren?

Mensch					**Maschine**	
+++	++	+	+/–	–	––	–––

Warum?
____
____
____
____
____

13.) Nachdem Sie nun SmartKom getestet haben, hat sich Ihre Einstellung gegenüber der Erkennung Ihrer Emotionen geändert? Warum?
____
_
____
_
____
____

14.) War die Aufgabe schwierig oder hatten Sie Schwierigkeiten in der Rolle zu bleiben?

O Ja O Nein

Warum?
____
____
____
___

15.) Das System SmartKom bietet die Möglichkeit, den Benutzer mithilfe verschiedener Techniken zu identifizieren. Würden Sie sich lieber mithilfe der Erkennung Ihrer Stimme oder Ihrer Unterschrift anmelden? Warum?

____
____
____
____
____

Teil III: Um den potentiellen Marktwert diese Systems besser einschätzen zu können, bitten wir Sie nun Fragen über die mögliche Benutzergruppe dieses Systems zu beantworten.

16.) Welche Personengruppe würde dieses System Ihrer Meinung nach am ehesten benutzen?

	benutzen					**nicht benutzen**	
	+++	++	+	+/–	–	––	–––
Leitende Angestellte							
Unternehmer							
Angestellte							
Freiberufler							
Arbeitslose							
Hausfrauen/–männer							
Kinder/ Jugendliche							
Studenten							
Senioren							
Jedermann							
Für mich							

17.) Wie sehr gefallen Ihnen folgende Aspekte von SmartKom:

	gefiel mir sehr					**gefiel mir gar nicht**	
	+++	++	+	+/–	–	––	–––
Displayeingabe (allgemein)							
Displayausgabe (allgemein)							
Spracheingabe (allgemein)							
Sprachausgabe (allgemein)							
Die Möglichkeit mit dem System „menschlich“ zu interagieren							
Das Design insgesamt							
SmartKom reagiert auf meine Emotionen							

18.) Was störte Sie am System SmartKom:

19.) Wieviel würden Sie für ein System dieser Art zahlen? (Angabe in Euro)

20.) Nennen Sie bitte eine von SmartKom **sehr schlecht** durchgeführte Aktion:

21.) Nennen Sie bitte eine von SmartKom **sehr gut** durchgeführte Aktion:

22.) Wie haben Sie sich bei der Benutzung des Systems SmartKom gefühlt?

O gut, ganz allgemein
O habe mich wohl gefühlt
O entspannt
O hat Spaß gemacht
O fasziniert
O sonstiges ____

O etwas unsicher
O angespannt
O ärgerlich
O ungeduldig
O ich kam mir beobachtet vor
O gelangweilt
O sonstiges ____

O Normal, nicht anders als sonst
O ganz ok
O sonstiges ____

Warum?
____
____
____
____
____

23.) Welche Elemente des Systems SmartKom vermitteln ein Gefühl von Innovation, Qualität, High–Tech, Spaß oder Nützlichkeit? **Bitte bewerten Sie die einzelnen Elemente mit einer Schulnote von 1 bis 6.**

(1= sehr gut bis 6= ungenügend).

	Innovation	Qualität	High–Tech	Spaß	Nützlichkeit
Displayeingabe (allgemein)					
Displayausgabe (allgemein)					
Spracheingabe (allgemein)					
Sprachausgabe (allgemein)					
Die Möglichkeit mit dem System „menschlich“ zu interagieren					
Das Design insgesamt					

24.) Welche dieser Elemente von SmartKom sollten sich an den Benutzer anpassen lassen?

____

____

____

___

25.) Wenn der SmartKom Assistent ein Auto wäre, welche Automarke wäre er dann? Warum?

____

____

____

____

_____

Teil IV: Nun zum Smartakus. Dies war die Persona, die Sie während der ganzen Nutzung in der oberen Ecke des Displays begleitet hat.

26.) Wie beurteilen Sie die **Kompetenz** von Smartakus?

Kompetent				**Nicht kompetent**			**Kann ich nicht sagen**
+++	++	+	+/–	–	––	–––	

Warum?

____
____
____
____
_____

27.) Wie natürlich wirkten die Aktionen der Persona auf Sie?

Natürlich				**Unnatürlich**			**Kann ich nicht sagen**
+++	++	+	+/–	–	––	–––	

28.) Wie gefiel Ihnen das Aussehen der Persona allgemein?

sehr gut						**gar nicht**	**Kann ich nicht sagen**
+++	++	+	+/–	–	––	–––	

29.) Hat Ihnen irgendetwas an der Persona besonders **GUT** gefallen? Wenn ja, was?

____
____
____
____

30.) Hat Ihnen irgendetwas an der Persona besonders **SCHLECHT** gefallen? Wenn ja, was?

____
____
____
____

Nun noch kurz einige wenige Fragen in eigener Sache:

31.) Hatten Sie an irgendeiner Stelle Schwierigkeiten mit diesem Fragebogen?
Wenn ja, wo?

____
____
____
____
_____

32.) Möchten Sie uns noch etwas mitteilen?

____
____
____
____
_____

SmartKom Mobil:

zu Frage 6: Bitte geben Sie an, wie sehr Sie folgenden Aussagen zustimmen:

	stimme zu					**stimme nicht zu**		**Nicht genutzt**
	+++	++	+	+/–	–	––	–––	
Die Überland–Karten waren zur Orientierung ausreichend								
Die Fußgänger–Karten waren zur Orientierung ausreichend								

SmartKom Public:

zu Frage 6: Bitte geben Sie an, wie sehr Sie folgenden Aussagen zustimmen:

	stimme zu					***stimme nicht zu***		***Nicht genutzt***
	+++	***++***	***+***	***+/–***	***–***	***––***	***–––***	
Die Gestikeingabe hat funktioniert								

SmartKom Home:

zu Frage 6: Bitte geben Sie an, wie sehr Sie folgenden Aussagen zustimmen:

	stimme zu					**stimme nicht zu**		**Nicht genutzt**
	+++	++	+	+/−	−	−−	−−−	
Das Telefonieren hat gut funktioniert								
Die Kinoauskunft entsprach meinen Wünschen								
Das TV−Programm ließ sich einfach aussuchen								
Die Programmierung des Videorekorders war einfach								
Die Stifteingabe hat gut funktioniert								

SmartKom Mobil:

zu Frage 17.) Wie sehr gefallen Ihnen folgende Aspekte von SmartKom:

	gefiel mir sehr					**gefiel mir gar nicht**		**Nicht genutzt**
	+++	++	+	+/−	−	−−	−−−	
Stifteingabe								
Identifikation über Stimmerkennung								
Identifikation über Unterschrift−erkennung								
Das Aussehen der Überland−Karten								
Das Aussehen der Stadt−Karten								

SmartKom Public:

zu Frage 23.) Welche Elemente des Systems SmartKom vermitteln ein Gefühl von Innovation, Qualität, High–Tech, Spaß oder Nützlichkeit? **Bitte bewerten Sie die einzelnen Elemente mit einer Schulnote von 1 bis 6.** Wenn Sie ein Element von SmartKom nicht genutzt haben, bitte kreuzen Sie **NUR** das entsprechende Feld an, ohne Notengebung.

(1= sehr gut bis 6= ungenügend).

	Innovation	**Qualität**	**High–Tech**	**Spaß**	**Nützlichkeit**	**Nicht genutzt**
Gestikeingabe						
Emotions–erkennung						
Die Kinoauskunft						
Das Telefonieren						
Das Versenden eines Dokumentes						
Der Assistent Smartakus						
Identifikation über Unterschrift–erkennung						

SmartKom Home:

zu Frage 23.) Welche Elemente des Systems SmartKom vermitteln ein Gefühl von Innovation, Qualität, High–Tech, Spaß oder Nützlichkeit? **Bitte bewerten Sie die einzelnen Elemente mit einer Schulnote von 1 bis 6.** Wenn Sie ein Element von SmartKom nicht genutzt haben, bitte kreuzen Sie **NUR** das entsprechende Feld an, ohne Notengebung.

(1= sehr gut bis 6= ungenügend).

	Innovation	**Qualität**	**High–Tech**	**Spaß**	**Nützlichkeit**	**Nicht genutzt**
Die Versendung von email						
Identifikation mit Unterschrift						
Identifikation über Stimm–erkennung						
Das Telefonieren						
Stifteingabe						
Program–mierung des Videorekorders						
Die Auskunft über das Fernseh–programm						
Die Kinoauskunft						

SmartKom Mobil:

zu Frage 23.) Welche Elemente des Systems SmartKom vermitteln ein Gefühl von Innovation, Qualität, High−Tech, Spaß oder Nützlichkeit? **Bitte bewerten Sie die einzelnen Elemente mit einer Schulnote von 1 bis 6.** Wenn Sie ein Element von SmartKom nicht genutzt haben, bitte kreuzen Sie **NUR** das entsprechende Feld an, ohne Notengebung.

(1= sehr gut bis 6= ungenügend).

	Innovation	**Qualität**	**High−Tech**	**Spaß**	**Nützlichkeit**	**Nicht genutzt**
Das Aussehen der Straßen−Karten						
Die Routenplanung allgemein						
Die Planung des Fußweges						

SmartKom Home:

33.) In welcher Situation würden Sie SmartKom nutzen wollen?

	sehr oft					**gar nicht**	
	+++	++	+	+/–	–	––	–––
Zuhause							
Im Urlaub							

SmartKom Public

33.) In welcher Situation würden Sie SmartKom nutzen wollen?

	sehr oft					**gar nicht**	
	+++	++	+	+/–	–	––	–––
Im Urlaub							
Auf Dienstreise							
Freizeit							

SmartKom Mobil:

33.) In welcher Situation würden Sie SmartKom nutzen wollen?

	sehr oft					**gar nicht**	
	+++	++	+	+/–	–	––	–––
Auf Dienstreise							
Auf dem Weg von/zur Arbeit							
In der Freizeit							
Im Auto							

Vielen Dank für Ihre Teilnahme an diesem Experiment und der Befragung.

12 Anhang C: Zusammenfassung der Paradise–Funktionalitäten

Paradise: A Framework for Evaluating Spoken Dialogue Agents

Die folgenden Abschnitte geben eine Zusammenfassung aus [1], [2], [3], [4], [5], [6], [7] über die Funktionalität und die Verarbeitungsweise von PARADISE (PARAdigm for DIalogue System Evaluation (von DARPA Communicators).

Entkoppelung von Anforderungen aus dem Anwendungsfall und dem Avartarverhalten (dem Dialogverhalten der Maschine)

Bisher:

Evaluation der Dialogstrategien über „inappropriate utterance ratio, turn correction ratio, concept accuracy, implicit recovery, transaction success“: Robustheit des Agenten (Systems) ergibt sich über die Wertung der Teilergebnisse verschiedener Dialogbewertungen.

(Dialogstrategie: Wie wird ein Dialog vom System geplant? Z.B. systemgeführter oder usergeführter Dialog, gemischter Dialog, vom allgemeinen zum besonderen, Komplettvorschläge)

Evaluation mit vordefinierter Referenzantwort [7]: Die Systemantwort auf eine Benutzerfrage wird mit einem vordefinierten Schlüssel von minimalen und maximalen Referenzantworten verglichen. Die Performanz ist der Anteil der Antworten, die auf den Schlüssel zutreffen. Nachteil: Es wird nur eine Dialogstrategie zugelassen.

Beide Ansätze können nicht ohne eine Identifikation der die Performanz verändernden Faktoren generalisiert werden. Auch Faktoren wie Datengröße und Umweltfaktoren wie Hintergrundgeräusche beeinflussen die Performanz.

Lösung: PARADISE

Vergleich von Dialogstrategien durch die Entkoppelung der benötigten Anwendungsfall–Anforderungen von der Art der Bearbeitung via Dialog. Die Performanz ergibt sich als eine gewichtete Funktion des anwendungsfallbasierten Erfolgsmaßes und dialogbasierter Kostenmessung, wobei sich die Gewichte aus der Korrelation von Benutzerzufriedenheit mit der Performanz ergeben.

Evaluierung des Prozentsatzes korrekter Antworten bzgl. einer Referenzantwort, der Prozentsatz erfolgreicher Transaktionen oder vervollständigter Tasks, Prozentsatz von angemessenen und unangemessenen systemdirektiven und diagnostischen Äußerungen, User Satisfaction (über Attribute Value Matrix).

Messungen können sehr stark korrelliert sein und das wird bei anderen Evaluationen nicht berücksichtigt.

Paradise erlaubt Generalisierungen zwischen Messungen.

Vergleich von Dialogstrategien

Performanzmodell für Dialoge: Methoden aus der Entscheidungstheorie für die Kombination von einer ungleichartigen Menge von Performanzmeßwerten (user satisfaction, task success, dialogue cost) in eine einzige Performanzfunktion zur Evaluation.

Performanz = $\alpha * \mathrm{N}(\kappa) - \sum_{i=1}^{n} \omega_i * \mathrm{N}(c_i)$

,wobei α ein Gewicht auf κ (task success), die Kostenfunktionen c_i durch ω_i gewichtet werden und N ist eine Z–Wert Normalisierungsfunktion.

Task Success: gemessen an Kappa–Koeffizienten. Kappa wird berechnet aus einer **ATTRIBUTE VALUE MATRIX[2]:**

Eine Attribute Value Matrix (AVM) stellt verschiedene Tasks des Dialogs dar.

Die Informationen, die zwischen System und Benutzer ausgetauscht werden, werden als eine Reihe von geordneten Attribut–Paaren und deren möglichen Werten dargestellt. (Tabelle 1[3]: Informationsfluß wird für die Evaluation nicht verwendet).

Tabelle2: AVM = Key für ein definiertes Szenario. Alle ***korrekten*** Durchführungen eines Szenarios hätten dieselbe Attributbelegung wie der Key.

AVM und TASK SUCCESS: <u>Konfusionsmatrix</u>

Der Erfolg einer Task (nicht Anwendungsfall wie in SK sondern (abstrakte) Aufgabenstellung) für einen ganzen Dialog (oder Subdialog) wird dadurch gemessen, wie gut das System und der Benutzer die Informationsanforderungen eines bestimmten Tasks erreichen. Dieselbe Task wird in mehreren Dialogen durchgeführt. Die Anwendungsfälle können dabei variieren.

Tabellen 3 & 4 zeigen zwei hypothetische *Konfusionsmatrizen*, die durch die Evaluation von 100 Dialogen mit jeweils zwei Zug–Stundenplan–Systeme A & B erstellt wurden.

Die Spalten zeigen den Key und spezifizieren dabei welche Informationswerte für ein bestimmtes Szenario der Benutzer und das System austauschen sollen. (Die Addierungen der jeweiligen Spalten in beiden Tabellen sind gleich => die Benutzer der beiden Systeme sollten dieselben Anwendungsfälle durchführen).

Die Zeilen stellen die vom Dialogkorpus gesammelten Daten dar. Sie zeigen,

---

[2] Zusammenfassung von [8]

[3] Alle Tabellen und Matrizen sind Originaldarstellungen aus [4], d.h. die Nummerierung der Tabellen und Dialoge enspricht der Originalnummerierung in [4].

welche Attributwerte eigentlich zwischen System und Benutzer vermittelt wurden. (Der Dialog muß bzgl. der Attribute getaggt werden (automatisch oder manuell).)

attribute	possible values	information flow
depart-city (DC)	Milano, Roma, Torino, Trento	to agent
arrival-city (AC)	Milano, Roma, Torino, Trento	to agent
depart-range (DR)	morning,evening	to agent
depart-time (DT)	6am,8am,6pm,8pm	to user

Table 1: Attribute value matrix, simplified train timetable domain

attribute	actual value
depart-city	Torino
arrival-city	Milano
depart-range	evening
depart-time	8pm

Table 2: Attribute value matrix instantiation, scenario key for Dialogues 1 and 2

Dialog 2

A1:	Hello, This is Train Enquiry Service.	DC,AC,DR,DT
	Please speak after the tone.	DC,AC,DR,DT
	Which information do you need?	DC,AC,DR,DT
U1:	I want to go from Torino to Milano.	DC,AC
A2:	Do you want to go from Trento to Milano?	DC,AC
	Yes or No?	DC,AC
U2:	No.	DC,AC
A3:	Do you want to leave from Trento?	DC
	Yes or No?	DC
U3:	No.	DC
A4:	Where do you want to leave from?	DC
U4:	I want to leave from Torino.	DC
A5:	Do you want to leave from Torino?	DC
	Yes or No?	DC
U5:	Yes.	DC
A6:	Do you want to go to Milano?	AC
U6:	Yes.	AC
A7:	At which time do you want to leave?	DR
U7:	I want to travel in the evening.	DR
A8:	Do you want to leave between 6 and 9 p.m.?	DR
	Yes or No?	DR
U8:	Yes.	DR
A9:	There is a train leaving at 8:00 p.m.	DT
	...	

Dialog 3

B1:	Hello, This is Train Enquiry Service.	DC,AC,DR,DT
	Please speak after the tone.	DC,AC,DR,DT
	Which information do you need?	DC,AC,DR,DT
U1:	I want to travel from Torino to Milano.	DC,AC
B2:	At which time do you want to leave from Merano to Milano?	DC,AC,DR
U2:	No, I want to leave from Torino in the evening.	DC, DR
B3:	Do you want to leave from Torino between 6 and 11 p.m?	DC, DR
	Please answer Yes or No.	DC, DR
U3:	Yes.	DC, DR
B4:	A train leaves at 8 p.m.	DT
	...	

	KEY													
	DEPART-CITY				ARRIVAL-CITY				DEPART-RANGE		DEPART-TIME			
DATA	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14
v1	**16**		1		4				3	2				
v2	1	**20**	1			3								
v3	5	1	**9**	4	2		4	2						
v4	1	2	6	**6**			2	3						
v5	4				**15**				2	3				
v6	1	6				**19**								
v7			5	2	1	1	**15**	4						
v8		1	3	3	1	2	9	**11**						
v9	2				2				**39**	10				
v10									6	**35**				
v11											**20**	5	5	4
v12												**10**	5	5
v13											5	5	**10**	5
v14												5	5	**11**
sum	30	30	25	15	25	25	30	20	50	50	25	25	25	25

Table 4: Confusion matrix, Agent B

user	task	agent state/strategy choices	US	κ	ET	STs	TOs	MRS	......
1	1	S_1:$A_{1,1}$, S_2:$A_{2,1}$, ...	24	1.0	220	18	15	.90	......
1	2	S_1:$A_{1,1}$, S_2:$A_{2,1}$, ...	27	.80	180	24	16	.79	......
1	3	....							......
......									......
48	3	....							......

Table 7: Form of Logged Information about Strategy Choices and Performance Metrics for a PARADISE Case Study: US = User Satisfaction, ET = Elapsed Time, Sts = System Turns, TOs = Timeout Prompts, MRS = Mean Recognition Score

Berechnung des Kappa–Koeffizienten aus der AVM:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

, wobei P(A) der Anteil der Übereinstimmungen der AVM–Szenarioschlüssel und der aktuellen AVMs ist und P(E) ist der Anteil der AVMs für die Dialoge.

Je unimodaler desto kleiner P(E).

Gibt es keine Übereinstimmungen wie erwartet ist $\kappa = 0$.

Bei totaler Übereinstimmung ist $\kappa = 1$.

κ ist anderen Maßeinheiten wie transaction success, concept accuracy und prozentualer Übereinstimmung überlegen, weil es die systematische Komplexität des Anwendungsfalls durch eine mögliche Korrektur des Dialogverlaufs berücksichtigt.

Die erwartete Übereinstimmung zwischen Daten und Schlüssel kann über die

Verteilung der Werte in den Schlüsseln abgeschätzt werden:

$$P(E) = \sum_{i=1}^{n} (t_i/T)^2$$

wobei t_i die Summe der Häufigkeiten in Spalte i der Matrix M ist und T die Summe der Häufigkeiten in M ($t_1 + ... + t_n$).

P(A), die aktuelle Übereinstimmung zwischen Daten und Schlüssel berechnet sich immer aus der AVM:

$$P(A) = \frac{\sum_{i=1}^{n} M(i,i)}{T}.$$

Berechnung der Kostenfunktionen c_i:

Berechnung der Kosten von qualitativen Meßwerten (z.B. No. of repair utterances).

c_i ergibt sich in der Regel aus der einfachen Aufsummierung der relevanten Attribute für den entsprechenden Meßwert.

Normalisierung über die Kosten:

Die Normalisierungsfunktion wird zum Ausgleich der Werte Kappa und der Kostenfunktionen verwendet, sie ergibt sich zu:

$$N(x) = \frac{x - \bar{x}}{\sigma_x}$$, mit σ_x Standardabweichung für *x und* $\bar{x}$ *Mittelwert*.

<u>Berechnung der Gewichtungen</u>: Die Gewichte werden beide über multiple linear regression auf Kappa und c berechnet.

Generalisierung

Über die AVM kann jede Task, jede Domäne und jeder Dialog abgearbeitet werden.

Die Attribute und Tags müssen allerdings angepaßt werden.

Der Dialog muß respektive in seine Dialogakte aufgeteilt werden, auf die dann die Attribute gesetzt werden.

Abfolge der Paradiseverarbeitung

Aufgabe und Anwendungsfälle bestimmen

AVM

Experimente mit unterschiedlichen Dialogsystemen für die Aufgabe

Berechnung der US über Fragebogen

Berechnung der task success

Berechnung der Dialogkosten über Effizienzmaß und Qualitätsmaß

Performanzabschätzung über lineare Regression und Werten für US, TS, DC

Anwendung der Performanzfunktion auf experimentelle Performanzdaten, um Performanzabschätzungen unter Systemstrategien, Anwendungsfällen und anderen experimentellen Variablen zu machen.

„Two Case Studies“

Systemumgebung

allgemeine Plattform für Sprachdialogsysteme mit ASR (sprachunabh. HMM),

Audioserver für Aufnahmen,

TTS (konkatenative Diphonsynthese) neben Stimmaufnahmen,

Telefoninterface,

Modul f. Applikation spez. Funktionen und Module für die Spezifizierung des Dialogmanagers und die Applikationsgrammatiken.

Möglichkeit für Barge−In

Dynamische Grammatikgenerierung und –laden (variables Vokabular im Task)

Dialogmanager (Automat), Knoten Systemantworten in Templates und Variablen (einleitend, Help auf Kommando, Ablehnung wenn Konfidenz d. Erkenners zu niedrig, timeout bei fehlendem Input)

Evaluationskriterien für beide Fallstudien

Aufnahmen zur teilweisen Verschriftung

qualitative Kostenmeßwerte:

completed (Task)

Barge in (Manuell gelabelt)

Dialogverhalten des Systems:

Timeout Prompts

ASR Rejections (Zurückweisung der Eingabe)

Help Requests

Mittelwert der Erkennungsrate (Mittlere Konzeptakkuratheit ACC manuell gelabelt)

Effizienz:

Elapsed Time (Dialogdauer)

No. Systemturns

No. Userturns

Kappa (aus erhaltener Information => AVM)

Verschiedene Dialogstrategien sollen getestet werden

ELVIS (email voice interactive system)

Allgemeines

Zugang zu Autor und Subject für Sammeladressen und Auswählen von Nachrichten.

systeminitiative vs. mixed initiative Dialogführung

größere Dialogeffizienz durch Inhaltsattribute (Sender, Subjekt)

3 Tasks mit je 2 Subtasks (Meetingzeit und Treffpunkt, Telefonnummer)

User Satisfaction (Fragen zur Bestimmung):

Was Elvis easy to understand in this conversation (TTS Performanz)

In this conversation, did Elvis understand what you said? (ASR Performanz)

In this conversation, was it easy to find the message you wanted?(Task Ease)

Was the pace of interaction with Elvis appropriate in this conversation (Interaction Pace != System Response!)

In this conversation, did you know what you could say at each point of the dialogue (User Expertise)

How often was Elvis sluggish and slow to reply to you in this conversation (System Response)

Did Elvis work the way you expected him to in this conversation(Expected Behaviour)

In this conversation, how did Elfiss voice interface compare to the touch–tone interface to voice mail(Comparable Interface gibts nicht in SK!)

From your current experience with using Elvis to get your email, do you think you'd use Elvis regularly to access your mail when you are away from your desk (Future Use)

– Befragung ueber 5wertige Skala (Häufigkeit) bzw. 3wertig (yes,no,maybe)

=> Zufriedenheit über Dialog. Berechnung der US (kumulativ): pro Dialog durch Aufsummierung der Multiplechoicefragen!!!!

Evaluation

48 User, 144 Dialoge, 6481 Turns, durchschnitt. Dialoglänge 5 Minuten,

Kappa 0.82,

Elapsed time 315 sec.

P(E) = 0.5,

P(A)= 0.91

$\alpha = 0.40$

$\omega = 0.78$

Performanz von SI > MI, MI schwerer zu lernen.

Erkennerperformanz mehr in Richtung US.

Nach 108 Trainingssessions ist U(systeminitiativ) auf 0.39 und U(mixed initiative) bei −0.43 geschätzt.

TOOT (online Amtrak Planungsinfo ueber Tel.)

Allgemeines

Bearbeitung von gespeicherten Tabellen und für User−Untertabellen wortwörtliche vs kooperativer Dialogstrategie für displaylose Infodarstellung

=> Hypothese: Performanz der Strategie ist abhaengig von der Task.

4 Tasks in Folge (Szenario: Zugverbindung mit best. Anforderungen suchen)

Evaluation

4 Tasks pro 12 User insg. 48 Dialoge in 1344 turns.

Kappa 0.81,

Dialogdauer 267 sec.

P(E) = 0.05

P(A) = 0.82.

komplexere Task als ELVIS (schwieriger AVM keys zu erraten)

Vergleich der Evaluationen

Kappawerte indizieren, daß beide Systeme bzgl. task success vergleichbar sind

Task completion perception und Erkennungsrate und BargeIn sind signifikant.

keine Effizienzmaße sind signifikante Voraussagen für die Performanz.

Effizienz ist eines der Hauptgewichte der Systemperformanz.

wörtliche und kooperative Dialogstrategien sind in ihrer mittl. Performanz nicht signifikant verschieden.

Verglichen mit ELVIS sind COMP, MRS, und ET signifikant Kompromiss zwischen Erkennerperformanz und Effizienz (s. auch Danieli und Gerbino)

Performanzangleichung:

Kombination experimenteller Daten über versch. Agenten inkl. Performanzfunktion.

Stochastische Optimalisierungsalgorithmen für automatische Verbesserung d.

Systems (genetische Algorithmen(Goldberg, Russell und Norvig), Q−lernen (Watkins), TD−Lernen (Tesauro), Adaptives Dynamisches Programmieren (Bellman, Sutton, Russell und Norvig, Barto)

Anwenden der Performanzfunktion auf den Dialog

Jeder Dialog hat eine Performanzbewertung.

Wenn versch. Auswahlmöglichkeiten für das System bzgl. Antwortstrategien: System kann optimale Lsg. finden.

Lernen online oder offline

Performanz als "Fitness"−Funktion, Basis Maximum Expected Utility

Voraussagen machen:

– Wenn eine Verbesserung der mittl. Erkennungsrate (ELVIS + TOOT) erreicht wird, kann auch die Performanz verbessert werden (z.B. durch andernen Erkenner)

– Verbesserung auch durch Dialogverhalten des Systems (z.B. systeminitiatives ELVIS, vertauschen der Prompts => einfacher zu verstehende Usereingaben)

=> adaptive Versionen von ELVIS und TOOT.

– Klassifizierer für problematische Dialoge via schlechter SE: future work.

Zusammenfassung:

– Genau festgelegte Arbeitsschritte

Unterstützung von Vergleichen zwischen Dialogstrategien,Task−Darstellung.

– Task success measure: Normalisiert für Taskkomplexität =>Basis für den Vergleich verschiedener Tasks.

– Performanzevaluation auf jeder Dialogebene.

– Kombination von Erfolgsmaßen (objektive und subjektive Kosten), Spezifizierung, der relativen Beiträge quantifiziert auf gesamte Performanz.

– Integration von Transaction success,, concept accuracy, multiple cost measures, US

– Task based success läßt keine besseren Möglichkeiten zu.(momentan)

– kein Unterschied zwischen falscher und nicht vorhandener Lösung

– Performanz als lineare Kombination von TS und DK => zusätzl. Fkt. benötigt.

Probleme bei der Evaluation mit PARADISE

Ergonomische Evaluation

Viele in PARADISE verwendete Bewertungskriterien können nicht oder nur eingeschränkt abgefragt werden, da SmartKom nicht mit unimodalen Dialogsystemen verglichen werden kann.

Daher müssen einige Fragen aus dem Paradise−Fragenkatalog komplett weggelassen bzw. ersetzt werden. Das sind im wesentlichen folgende:

„Comparable Interface“: Es gibt keine multimodalen Standardsysteme, mit denen verglichen werden kann.

„Interaction Pace vs. System Response“: Wie kann eine VP objektiv einen Dialog auswerten, bei dem sie selbst mitwirkt?

Die Skalen zu den einzelnen Fragen sind nicht homogen (3 vs. 5 Wertungen) und zudem nicht zuweisbar.

Außerdem sind einige Berechnungen fraglich: Warum werden zur Bewertung der User satisfaction Absolutwerte (Summe ueber die Fragen, deren Antworten als Mittelwerte über alle VPn geht) anstatt Mittelwerte genommen? Wie können diese interpretiert werden?

Technische Evaluation

Auch wenn es bei der Berechnung der „Interaction Pace vs. System Response“ nicht zu Kollisionen in der Evaluation kommt, da meßbar objektiv zwischen Dauer des Gesamtdialogs und der mittleren Dauer der Systemantwortzeit klar unterschieden werden kann, kommt es auch hier zu einigen grundsätzlichen Problemen.

Zu nennen ist hier insbesondere die Verwendung von Absolutwerten bei der Kostenfunktion, die nicht interpretierbar wird, die Berechnung der Performanzgewichte über MLR sowie die Definition von Referenz–Keys für die AVMs.

Probleme bei der Portierung der Evaluationsrichtlinien von PARADISE auf multimodale Dialogsysteme

Die Probleme, die sich bei der Portierung der Evaluationsrichtlinien von PARADISE auf SmartKom ergeben, liegen natürlich klar auf der Hand:

Wie geht man mit multimodalen Eingabe– und Ausgabemedien um?

Wie verhält es sich mit der Evaluation der drei Szenarien? Wird hier ebenso wie über die Anwendungsfälle normalisiert? Wenn ja, wie? Wenn nein, haben wir dann eigentlich 3 Systeme?

Bei der Berechnung der AVMs für SmartKom entstehen multidimensionale Matrizen. Wie kann man hier einen optimalen Dialog rausfinden?

Wie geht man mit der so entstehenden Streuung um?

Literatur

[1] M.A. Walker, et al.: DARPA Communicator Dialog Travel Planning Systems: The JKune 2000 Data Collection, LREC 2000.

[2] M. Rahim et al.: VOICE–IF: a mixed–initiative spoken dialogue system for AT&T conference services. EUROSPEECH 2001

[3] M. Rogati et al.: Training a Sentence Planner for Spoken Dialog: The Impact

of Syntactic and Planning Features. EUROSPEECH 2001

[4] M.A. Walker et al.: Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies,In Computer Speech and Language, Vol. 12, No. 3, 1998.

[5] M.A. Walker, D.J. Litman, C.A. Kamm, A. Abella: PARADISE: A Framework for Evaluation Spoken Dialogue Agents. Annnual Meeting of the Association of Computational Linguistics. ACL 1997.

[6] U. Andenfilger: Some Remarks about the Validation of Information Systems Development. In Interdisciplinary Foundation of Systems Design and Evaluation. Seminar Report 97 of International Conference and Research Center for Computer Science, edited by R. Keil–Slawik, Saarbrücken Schloss Dagstuhl, University of Saarland.

[7] L. Hirschmann, D.A. Dahl, D.P. McKay, L.M. Norton, M.C. Linebarger: Beyond class A: A proposal for automatic evaluation of discourse. In Proceedings of the Speech and Natural Language Workshop,1990.

[8] V. Penide: ATTRIBUTE VALUE MATRIX. Detaillierte Beschreibung zur Berechnung der AVMs. ~smart/EVAL/AVM/AVM–paradise.sdw

Anhang

Zusammenfassung der Beschlüsse vom 29.01.02

<u>Die publizierte Form von PARADISE ist für SmartKom nicht anwendbar =></u>

Die SK–End–to–End–Evaluation darf NICHT PARADISE genannt werden! (D.h. „Evaluation in Anlehnung an PARADISE“, besser EIGENER Name)

Eventuell können dennoch Teile in adaptiver Form daraus verwendet werden

Berechnung der Einzelkosten (wie schon im ersten Eval–Entwurf geplant)

Keine Vorgabe der Keys in SK (keine Slotorientierung, dynamische Definition der Keys, finiter Automat) => keine vernünftige Berechnung der AVM

Alle potentiellen Einträge sind Key–Beteiligte => bei TV hat Key: Film eine 2000x2000 Matrix

Durch VP Instruktion die Matrix einschränken

Keys aus Modul (Dialogmanager?)

Definition der Keys über ein Range der tatsächlich in einer Stichprobe aus den Aufnahmen vorgekommenen Werte

Gruppierung der VP nach Anwendungsfall, Evaluierer legen Benutzerkeys fest.

=> Task Success (Kappa) an WOZ testen => neue Strategie zur Berechnung der Keys in der AVM und zur Berechnung von P(E) (Uli testet)

Keytagger über alle Eingaben

Normierung ok, da über MLR die Berechnung der Kosten bzw. keinen Unterschied macht, solange nicht gemischt wird.

Momentan nicht lösbar bei der Evaluierung nach Paradise:

Wie soll die Multimodalität vernünftig bewertet werden? (Vierfachansicht?)

Wie können die Erkennerleistungen vernünftig gewichtet werden?

Wie bewertet man einen unvollständig durchgeführten Task? Über die AVM würde trotzdem ein geringer Task Success eingehen, obwohl der Task zu KEINEM Erfolg führt.

Wie wird der Task success bewertet, wenn der User für das Nicht–Gelingen verantwortlich ist (aus der Rolle fällt)?

Änderungen an der bisherigen Evaluationsdefinition:

Anstelle einer MLR Berechnung über t–Test auf jede Kostenfunktion in Abhängigkeit der entsprechenden Fragebogenfrage sinnvoller =>

Anpassung des Fragebogens: Aufsplitten allgemeiner Fragen z.B. zur Erkennung, Darstellung der Ergebnisse durch das System.

Zusammenfassung der Beschlüsse vom 19.02.02

Die nachfolgenden Lösungsvorschläge beziehen sich auf die in 5.1 genannten Probleme:

ad 1) Die SmartKom−Evaluation nach erweitertem Paradise heißt:

PROMISE − A Procedure for Multimodal Interactive System Evaluation

ad 2.1) Siehe: ~smart/EVAL/TE/TE_Toolspezifikation.sdw bzw.

~smart/EVAL/TE/paradise−smartkom.sdw

ad 2.2) Berechnung einer „AVM“:

Gegeben: Kooperativität s. ad 3.4

Gegeben: Task completed s. ad 3.3

Definition von „Minimalpaaren“ und Keywordspotting

Für jeden Anwendungsfall werden die Kombinationen potentieller Keys festgelegt, mit denen eine eindeutige Task success gegeben ist.

Z.B.: TV

Die Bestimmung des Abendprogramms ergibt sich durch die Auswahl eines Films. Die Task ist erfolgreich, wenn mindestens

Der Filmtitel genannt wird

Kanal und Uhrzeit genannt werden

...

Es gibt dann nur die Werte 1 bzw. 0 in der Bewertung.

Falls es keine „Minimalpaare“ gibt, kann ein Task trotzdem erfolgreich sein, wenn es durch den User angezeigt wird (Keywordspotting)

Z.B.: TV

Ein User läßt sich zunächst eine Auswahl an Filmen geben, um später eine Entscheidung treffen zu können, es werden eindeutige Äußerungen gemacht, die auf die positive Beendigung eines Tasks deuten („ Schade, da ist nichts für mich dabei. Was kommt denn im Kino?“ (s. Demonstratordialog. Für das TV−Programm kann SK ja nichts) vs. „Ich will das Fernsehprogramm für heute Abend und nicht für morgen! Na, dann zeig mir mal, ob du mit dem Kino besser zurechtkommst“)

Festlegung des Keywords in der Versuchsanleitung

Gewichtung von Kappa (Task Success) noch unklar.

ad 3.3) Ein Task ist vollständig, wenn es eine aktive Beendigung des Tasks durch den User gibt: Anwendungsfall wechselt.

Alle unvollständigen Tasks gelten als abgebrochen (entweder aufgrund von Unkooperativität s. 3.4 oder falls 2.1 nicht erfolgreich und ebenfalls 2.2 nicht erfolgreich) und werden extra gewichtet (für 2.1/2.2; wie, ist noch unklar).

Noch zu klären: Aufnahmeende = abgebrochener Task???

ad 3.4) Durch die Versuchsanleitung werden unkooperative User herausgefiltert:

Fällt ein User aus der Rolle oder verhält sich unkooperativ (z.B. ständig Schachtelsätze, trotz Hinweis des Systems: „Bitte verwenden Sie kurze Sätze“), wird die Aufnahme durch die VL abgebrochen und gegebenenfalls wiederholt.

Motivation für VPs zur Kooperativität: Geld gibt’s nur, bei „Task success“ (bzw. falls das System der Grund für fehlende Task success)

Anzahl der Abbrüche hier sind nicht Teil der Performanzbewertung.

ad 4) Fragebogen wird angepaßt.

Noch zu klären:

3.1

3.2

Aufnahmeende = abgebrochener Task???

Keywords definieren

Kappa−Gewichtung

Gewichtung unvollständiger Tasks bei Kooperativität