SmartWeb Evaluierungsworkshop Protokoll

Moritz Kaiser
IPSK
Ludwig-Maximilians-Universität München
ariser@phonetik.uni-muenchen.de

6. Juli 2006

Der Evaluierungsworkshop befasste sich mit der Evaluierung, die vom IPSK durchgeführt werden soll. Die Wünsche der Teilnehmer wurden gesammelt und versucht einen Konsens über die Evaluierungsprozedur zu finden. Hauptthema war die Zielsetzung der Evaluierung. Aufbauend darauf wurden Art und Umfang der Evaluierung beschlossen.

Inhaltsverzeichnis

Protokoll **3**

1 Begriffsklärungen **4**
1.1 Einteilung 4
1.1.1 Datenquelle 4
1.1.2 Ablauf 4
1.1.3 Anwendung 4
1.2 Benutzerseite 5
1.3 Technik und Analyse 5

2 Erfahrungen mit V 0.4 **5**

3 Vorschlag LMU **6**

4 Diskussion und Beschlüsse **7**

A Entscheidungen **12**

B Nachträgliche Entscheidungen nach Rückfragen **13**

Protokoll

Moderation Florian Schiel

Protokollant Moritz Kaiser

Anwesend Jürgen Vogel (EML)
Florian Schiel (LMU)
Hannes Mögele (LMU)
Moritz Kaiser (LMU)
Wiebke Johannsen (DTAG)
Bettina Säuberlich (IMS)
Daniel Sonntag (DFKI)
Jochen Leidner (Uni SB)
Ralf Engel (DFKI)
Ralf Decke (BMW)
Anton Batliner (FAU)

Sitzungsort IPSK, LMU, Schellingstr. 3, München

Datum 21. Juni 2006 10:00 – 16:15

Tagesordnung

1 Begriffsklärungen **4**
1.1 Einteilung 4
1.2 Benutzerseite 5
1.3 Technik und Analyse 5

2 Erfahrungen mit V 0.4 **5**

3 Vorschlag LMU **6**

4 Diskussion und Beschlüsse **7**

A Entscheidungen **12**

B Nachträgliche Entscheidungen nach Rückfragen **13**

1 Definitionen und Begriffe der Evaluierung

Florian Schiel stellt die verschiedenen Begriffe, Konzepte und Definitionen, die für eine Planung der Evaluierung notwendig sind, vor.

1.1 Einteilung der Arten der Evaluierung

1.1.1 Nach Datenquelle

- Subjektive Evaluierung bezeichnet die Beschreibungen und Einschätzungen der Versuchspersonen. Diese wird durch Befragung mittels genormter Fragebögen ermittelt und kann z.B. Einschätzungen der Systemeigenschaften liefern.
- Technische Evaluierung bezeichnet die objektive Erfassung von Systemparametern und deren Vergleich mit der Realität. Darin enthalten sind z.B. Antwortzeiten einzelner Module, Fehlerhäufigkeit der Spracherkennung usw.

1.1.2 Nach langfristigem Ablauf

- Kurzzeitevaluierung beschreibt eine Vorgehensweise, bei der jede Versuchsperson zu einem Test mit 30 bis 60 Minuten Dauer eingeladen wird. Die Zahl der VPen ist daher relativ hoch. Bei dieser Art der Evaluierung kann die Verwendbarkeit durch Erstbenutzer beobachtet werden.
- Langzeitevaluierung beschreibt eine Vorgehensweise, die mehrere Sitzungen pro VP ansetzt. Es wird dabei unterschieden zwischen einer Trainingsphase, die meist nicht ausgewertet wird, und einer eigentliche Evaluierungsphase. Die Anzahl der VPen ist aufgrund des höheren Einzelaufwands niedriger.

1.1.3 Nach Anwendung

- Begleitende Evaluierung findet während des Entwicklungsprozesses statt. Sie soll vor Ende dessen die Verwendbarkeit bewerten und Verbesserungsvorschläge für nachfolgende Prototypen liefern. Der untersuchte Prototyp wird während der Evaluierung regelmäßig auf den jeweils aktuellen Stand gehoben.
- End-to-end Evaluierung findet nach Fertigstellung eines der letzten Prototypen statt. Dieser wird im Zustand dieser Release eingefroren. Ziel ist eine abschließende Bewertung, die erst einem Nachfolgeprojekt zu Gute kommen kann. Die Bewertung dient Projektträgern und Geldgebern zur abschließenden Kontrolle.

1.2 Festlegungen auf Benutzerseite

Diese betreffen in erster Linie die Aufgaben, mit denen der Benutzer konfrontiert wird, wirken sich aber auch auf das System aus.

- Use-Cases beschreiben jeweils eine Aufgabe oder einen Aufgabenkomplex, die von den VPen im Test zu lösen ist. Ein Use-Case bezieht sich in der Regel auf eine Domäne, in der eine oder mehrere zusammenhängende Informationen errungen werden sollen.
- Szenarien beschreiben die Umgebungen bzw. Einsatzorte des Systems. Im Gegensatz zum Use-Case steht nicht die zu findende Information im Vordergrund, sondern die Einflüsse von Umgebungsgeräuschen, Ablenkung etc. auf VP und System.

1.3 Technik der Datenerfassung und Analyse

- Technisches Setup ist die Konfiguration des Clients. Es beschreibt die aktiven Module, die auswertbaren Datenquellen, die geführten Log-Dateien.
- Monitoring erfasst die vom System unabhängigen Realdaten, die z.B. mit einer Videokamera von Versuchsperson und Umgebung erfasst werden.

2 Erfahrungen mit dem Prototyp V. 0.4 an der LMU

Es wurde mit vier verschiedenen Sprechern jeweils ein Probedurchlauf in ruhiger Umgebung durchgeführt. Der Probedurchlauf enthielt einen Großteil der von der Systemgruppe als ‚funktionierend' deklarierten Fragen.

- Einige wenige der ‚erlaubten' Inputs funktionierten gar nicht.
- Inputs außerhalb der Liste funktionieren i.d.R nicht.
- Die Antwortzeiten sind akzeptabel.
- UMTS als Trägermedium funktioniert, wurde allerdings nur ohne Kamera getestet.
- Oft werden entgegen der Ankündigung keine Medieninhalte geliefert.

Die Installation erwies sich als problemlos. Insbesondere erscheinen

- die Möglichkeit die Dienste auf zwei System zu verteilen
- die klar definierten Ports
- der übersichtliche Installationsvorgang
- und die Möglichkeit der Installation über Subversion

als sehr förderlich für eine weiter Evaluierung. Zusammenfassend lässt sich sagen, dass der Prototyp V. 0.4 für die Evaluierung nicht geeignet gewesen wäre, da das Anwendungsspektrum zu schmal war. Große Teile des Systems funktionieren aber gut und vor allem die Serverdienste scheinen relativ stabil zu laufen.

3 Vorschlag des IPSK zur Durchführung der Evaluierung

Der Vorschlag gründet sich auf Erfahrungen aus SmartKom und den ersten Tests mit dem System. Die Evaluierung soll in zwei Phasen geteilt werden:

- August bis Dezember 2006: Begleitende Evaluierung mit 10 VPen
- Januar bis September 2007: End-to-End Evaluierung mit mehr als 10 VPen

Es sollen ausschließlich Langzeitevaluierungen stattfinden mit fünf bis zehn Sitzungen pro VP. Diese Sitzungen sollen ca. 10 Minuten dauern. Als Use-Cases kommen in Frage:

- Fußball-WM
- Wetter
- Navigation
- freie Anfragen
- weitere Domänen, so vom Prototyp unterstützt.

Folgende Szenarien werden vorgeschlagen:

- ruhige Büroumgebung
- Park
- Straße
- Bahnhof

Der Ablauf einer Langzeitevaluierung soll zeitlich wie folgt strukturiert werden:

1. Eine Trainingssitzung unter optimalen Bedingungen (Büro)
2. Die Versuchsperson soll auf ausreichende Benutzerkompetenz geprüft werden, um sinnlose Evaluierungen auszuschließen.
3. Training evtl. mit weiteren Sitzungen
4. fünf bis zehn Sitzungen mit steigendem Schwierigkeitsgrad

Das technische Setup soll folgende Eingabemodalitäten ermöglichen:

- eingebautes Mikrophon
- Tastatur

- Stift

Weiterhin soll ein Mitschnitt der Gesichtskamera erreicht werden. Die Aufnahmen sollen sowohl über UMTS als auch über WLAN durchgeführt werden.

4 Diskussion und Beschlüsse zur Evaluierung

Bezüglich der Art der Evaluierung stimmt die Versammlung mit dem Vorschlag überein.

Beschluss: Beginn der Evaluierung als begleitende Evaluierung

> Es soll der maximale Nutzen aus der begleitenden Evaluierung gezogen werden, bevor die End-to-end Evaluierung beginnt.

Beschluss: End-to-End Evaluierung ja.

> Nach Abschluss der begleitenden Evaluierung soll der Prototyp eingefroren werden und dann eine abschließende Evaluierung entsprechend des Vorschlags stattfinden

Gegenüber dem Vorschlag votiert Anton für eine deutliche Reduzierung der Anzahl der Szenarien, um eine Zersplitterung zu vermeiden.

Beschluss: Drei Szenarien

> Es werden nur die Szenarien Büro, Café und Straße aufgenommen.

Beschluss: Eigenschaften der Szenarien

> Im Büro soll die Ablenkung des Benutzers niedrig sein, auf der Straße hoch. Lärmpegel ist selbsterklärend

Demgegenüber werden aufgrund der erweiterten Eigenschaften des Demonstrators 0.5 wesentlich umfangreichere Use-Cases definiert als im Vorschlag.

Beschluss: Use-Cases mit teiloffenen Domänen

> Es sollen Tests mit Personen, Orten und Bildern durchgeführt werden

Beschluss: Use-Cases mit geschlossenen Domänen

> Domänen sind Wetter, Fußball, Navigation, Verkehr, Kino und Webcams

Beschluss: Einführung der VPen in Domänen

> Die Benutzer müssen zu den Domänen hingeführt werden, ohne die Freiheit zu stark einzuschränken. Teils müssen die Domänen vorgegeben werden.

Über die Vorauswahl der Benutzer herrscht Uneinigkeit.

- ⊖ Eine Vorauswahl verhindert, dass die Benutzbarkeit des Systems an einem repräsentativen Querschnitt gemessen werden kann.

⊕ Die Vorauswahl verhindert, dass 10 Evaluierungssitzungen mit technisch nicht trainierbaren Personen verschlissen werden. Die Erkenntnisse aus solchen Sitzungen wären wertlos.

⊖ Die Vorauswahl wird immer subjektiv sein.

⊕ Es können Kriterien für die Vorauswahl aufgestellt werden.

Beschluss: Selektion ja oder nein?

> Eine Selektion muss zur Effizienzsteigerung durchgeführt werden.

Beschluss: Zeitpunkt der Selektion

> Eine Selektion wird nach einer oder zwei Probeaufnahmen und folgender Befragung durchgeführt.

Beschluss: Bedingung zur Selektion

> Die Abgelehnten VPen müssen genau so dokumentiert werden, wie die aktiven. Außerdem müssen die Ausschlussgründe dokumentiert werden.

Vermutlich kann nur das eingebaute Mikrophon genutzt werden. Dies steht allerdings im Widerspruch zum Korpus, der ja über Bluetooth aufgezeichnet worden ist. Ob andere Mikrophone verwendet werden können und sollen, kann nicht mit letzter Sicherheit geklärt werden.

Beschluss: Mikrophonie

> Mit Gerd Herzog muss geklärt werden, ob es beim eingebauten Mikro bleibt, oder ob doch noch andere Quellen implementiert werden

Weiterhin hat sich gezeigt, dass die Ausgabelautstärke des Clients mehr als ungenügend ist. In lauteren Umgebungen ist die Synthese unhörbar.

Beschluss: Lautsprecher

> Mit der Systemgruppe muss geklärt werden, ob sich nicht eine höhere Ausgabelautstärke erzielen lässt.

Das Protokollieren der verschiedenen Module ist offensichtlich nicht einheitlich gestaltet worden. Vermutlich reichen die Informationen in den Protokolldateien nicht an jeder Stelle für die Evaluierung aus.

Beschluss: Protokolldateien

> Von den Partnern müssen über den Weg der MoKo Mindestinformationen in den Protokolldateien erbeten werden.

Tastatur und Stifteingaben sollten laut Systemgruppe im Log zu finden sein, ebenso die Display und Syntheseausgaben.

Beschluss: Protokolldateien und Zeitstempel

> Zeitstempel in den Protokolldateien sind unumgänglich.

Für die Evaluierung des On-View-Erkenners ist eine Videokamera Richtung Benutzer nötig. Es ist zunächst nicht klar welches Endgerät nun dafür verwendet werden kann. Es wird die Frage aufgeworfen, ob Video überhaupt evaluiert werden muss.

⊖ Die On-View Klassifikation wird im Demonstrator überhaupt nicht verwendet. Das ist auch nicht geplant.

⊕ Die On-View Klassifikation könnte aber verwendet werden, weil sie eine hohe Spezifität liefert.

⊖ Es ist zu unsicher die On-View Klassifikation für die Aktivierung des Erkenners zu benutzen.

⊕ Die On-View Klassifikation könnte aber zur Steuerung der Synthese verwendet werden.

Beschluss: On-View

> On-View soll evaluiert werden und zwar anhand des ausgegebenen Videostreams mit Markern.

Es wird vermutet, dass die Bandbreite von UMTS für On-View nicht ausreicht. Es kommt also nur WLAN für die Evaluierung von On-View in Frage. Die Benutzerkamera des MDA Pro ist vermutlich nicht ansprechbar. Folgende Konfigurationen sind daher für die Video-Evaluierung denkbar:

- MDA Pro mit ‚umgedrehter' Haltung (unpraktisch)
- MDA Pro mit Benutzerkamera (Implementierung sehr unwahrscheinlich)
- MDA III mit Aufsteckkamera (Beschaffungsproblem)
- MDA Pro mit Spiegeloptik (evtl. unhandlich)

Die Lösung hängt zum einen von der Beschaffbarkeit eines MDA III bzw. von der Handhabbarkeit des Spiegelsystems ab.

Beschluss: MDA III

> Das IPSK muss versuchen einen MDA III mit Kamera bei den Partnern zu aquirieren.

Es stellt sich die Frage, ob ein Mehrbenutzerbetrieb, der ja vorgesehen ist, getestet werden soll. Die Anforderungen an die Hardware wären dabei relativ umfangreich und der Nutzen des Tests relativ zweifelhaft.

Beschluss: Mehrbenutzerbetrieb

> Es wird kein Mehrbenutzerbetrieb getestet, weil der Aufwand zu groß wäre

Für den Betrieb des Demonstrators V. 0.4 zeigt sich eine 95%ige Auslastung des Arbeitsspeichers des Servers. Da die späteren Versionen eine eher mehr und größere Komponenten enthalten werden, empfiehlt sich ein Ausbau des Arbeitsspeicher. Auch die

Prozessorleistung kann zu Engpässen führen. Die Prozesse können jedoch auf zwei Rechner verteilt werden.

Beschluss: Server

> Der Server wird auf $4 * 10^9$ Byte RAM aufgerüstet. Bei Bedarf wird ein zweiter Rechner verwendet

Einige Anmerkungen zum Server:

- Es existiert eine kritische Frage, die die Belastung des Servers maximiert: „Gegen wen spielte Frankreich bei der WM 1998?“Die Ursache dafür ist unbekannt, der Effekt eignet sich jedoch zum Performancetest.
- Subversion sollte verwendet werden, wenn begleitend evaluiert wird. Dadurch können Evaluierungs kritische Änderungen schneller eingepflegt werden.

Die Evaluierung der Spracherkennung ist eine der wichtigsten Teilaufgaben. Zur Evaluierung muss Zugriff auf die Audiodaten bestehen.

Beschluss: Spracherkennung

> Die Evaluierung der Spracherkennung erfordert das Speichern der Audiodaten und der passenden Zeitstempel.

Beschluss: Off-Talk

> Die Evaluierung der Off-Talk-Erkennung soll abhängig von der Inzidenz im Material erfolgen.

Von einigen Partnern wird gewünscht, auftretende Systeminitiativen zu protokollieren.

Beschluss: Systeminitiativen

> Die Aktivitäten des recommender systems sollen, sobald sie in Erscheinung treten aufgezeichnet und evaluiert werden.

Die Subjektive Evaluierung soll nach allgemeinem Konsens über die reine Befragung hinausgehen. Primäre Quelle bleiben jedoch Fragebögen.

Beschluss: Systeminitativen

> Die Systeminitiativen sollen vom Benutzer bewertet werden. Es soll beobachtet werden, wie der Nutzer darauf eingeht.

Beschluss: Fragebögen

> Fragebögen sollen den Nutzer zu Benutzbarkeit, Synthesequalität, Nützlichkeit von Links, Medien und Akkuratheit der Informationen befragen.

Beschluss: Beobachtungen

> Die Nutzung der GUI und der Links des Benutzers soll beobachtet werden.

Es besteht der Vorschlag, SOT zur subjektiven Evaluierung heranzuziehen. Dazu muss aber SOT im Audiomaterial eindeutig identifiziert werden. Dies ist eigentlich nur mit PTT möglich.

Beschluss: PTT

> Die Systemgruppe soll entscheiden, ob es möglich ist, einen expliziten PTT-Knopf am Gerät einzuführen

Beschluss: SOT

> SOT soll evaluiert werden, wenn PTT möglich ist.

Beschluss: Audioquellen des Monitoring

> Als Audioquelle soll lediglich ein Ansteckmikrofon dienen. Dieses soll über die Tonspur der Videoaufnahme mitgeschnitten werden.

Es wird besprochen, ob mit der Videokamera ein Mitschnitt des Geschehens auf dem Display möglich ist. Das wird jedoch als unwahrscheinlich eingeschätzt.

Beschluss: Methode des Videomitschnitts

> Der Versuchsleiter filmt die Versuchsperson mit Fokus auf die Handlungen der VP und den Einfluss der Umgebung.

Beschluss: Auswertungswerkzeug für Protokolldateien

> Um die Auswertung der Protokolldateien zu erleichtern soll die Systemgruppe gebeten werden, ein Tool bereit zu stellen.

Beschluss: Evaluierung der Motorradumgebung

> Diese wird durch BMW selbst durchgeführt und soll ähnliche Daten erfassen, wie die Handheldaufnahme. Video ist überflüssig.

Beschluss: Training der Benutzer

> Um das Training zu erleichtern sollen die Benutzer mittels eines Demovideos und einer Flashdemo durch den VL an das System herangeführt werden.

Es bleibt offen, wie intensiv man die VPen in das System einführen soll und wie stark die Eigeninitiative dabei behindert wird.

A Entscheidungen

Art der Evaluierung
Beginn der Evaluierung als begleitende Evaluierung . . . 7
End-to-End Evaluierung ja. . . . 7
Definition der Szenarien und Use-Cases
Drei Szenarien . . . 7
Eigenschaften der Szenarien . . . 7
Use-Cases mit teiloffenen Domänen . . . 7
Use-Cases mit geschlossenen Domänen . . . 7
Einführung der VPen in Domänen . . . 7
Aussortieren unbrauchbarer VPen
Selektion ja oder nein? . . . 8
Zeitpunkt der Selektion . . . 8
Bedingung zur Selektion . . . 8
Rückfragen, die vor weiteren Entscheidungen anstehen
Mikrophonie . . . 8
Technische Evaluierung
Lautsprecher . . . 8
Protokolldateien . . . 8
Protokolldateien und Zeitstempel . . . 8
On-View . . . 9
MDA III . . . 9
Notwendige Infrastruktur
Mehrbenutzerbetrieb . . . 9
Server . . . 10
Spracherkennung . . . 10
Off-Talk . . . 10
Systeminitiativen . . . 10
Subjektive Evaluierung
Systeminitativen . . . 10
Fragebögen . . . 10
Beobachtungen . . . 10
PTT . . . 11
SOT . . . 11
Monitoring der Versuche
Audioquellen des Monitoring . . . 11
Methode des Videomitschnitts . . . 11
Auswertungswerkzeug für Protokolldateien . . . 11
Sonstige Entscheidungen
Evaluierung der Motorradumgebung . . . 11
Training der Benutzer . . . 11

B Nachträgliche Entscheidungen nach Rückfragen

- Die Rechner des IPSK können nicht sinnvoll bis vier GiByte aufgerüstet werden, da es sich nicht um Vierundsechzigbitsysteme handelt. Eine Aufrüstung auf drei GiByte wird durchgeführt, zusätzlich werden SATA-Festplatten nachgerüstet um u.a. das System von Wartezeiten beim Schreiben der Logdateien zu entlasten. Es werden entgegen dem Beschluss sofort zwei Rechner eingerichtet.
- Eine PTT-Implementierung ist bis auf Weiteres ausgesetzt. Grund dafür ist, dass damit ein anderes System evaluiert würde, als tatsächlich eingesetzt. Kommt die begleitende Evaluierung zum Schluss, dass das Berührfeld eine schlechte Bedienbarkeit zeigt, kann das gesamte System auf PTT umgestellt werden.
- Vergessen wurde auf dem Eval-Workshop die OOV-Erkennung. Geloggt und eventuell evaluiert werden sollte die phonetische Kette.
- Die Vermutung vom Workshop, dass On/Off/View/Talk-Ergebnisse nicht weiter verwendet werden, ist wohl falsch. Die Ergebnisse sollen sogar zur Ablehnung von Anfragen führen: "Haben Sie mich gemeint?". Dies macht eine Evaluierung deutlich interessanter.