Evaluierung einer intelligenten Hausumgebung durch Modellierung von Benutzerverhalten

Sebastian Möller, Klaus-Peter Engelbrecht

Deutsche Telekom Labs, TU Berlin, Deutschland, Email: [sebastian.moeller;klaus-peter.engelbrecht]@telekom.de

Einleitung

Je bessere Möglichkeiten bestehen, Umgebungen durch Einsatz von Sprachtechnologie „intelligent" zu gestalten, desto stärker steigt der Bedarf an einer schnellen und ökonomischen Evaluierung. Üblicherweise umfasst eine adäquate Evaluierung zwei Aspekte: Die Überprüfung der Leistung der beteiligten Systemkomponenten (z.B. Spracherkennung, Sprachverstehen, Dialogführung und Sprachausgabe), sowie die Quantifizierung verschiedener Qualitätsaspekte aus Benutzersicht, wie bspw. der Effizienz, des Komforts, der Gebrauchstauglichkeit sowie der Akzeptanz. Da sich Qualität als Ergebnis eines Wahrnehmungs- und Beurteilungsprozesses ergibt, bedarf es zur Messung der oben genannten Qualitätsaspekte i. Allg. kontrollierter (Labor-) Experimente mit Versuchspersonen.

In diesem Beitrag soll ein neuer Weg beschritten werden, um Qualität und Gebrauchstauglichkeit mit nur minimalem Einsatz von Versuchspersonen schon in der Designphase zu quantifizieren. Dazu wurde Benutzerverhalten im Umgang mit einer Sprachsteuerung für Hausgeräte analysiert und „fehlerhaftes" Verhalten, welches zu einer unnötigen Verlängerung der Interaktion führt, phenotypisch klassifiziert. Es wurden fünf Klassen von Fehlern extrahiert, welche nun mit Hilfe eines Benutzermodells explizit generiert werden sollen. Dadurch kann das Verhalten des Systems bei nicht optimalem Benutzerverhalten quantifiziert und Vorhersagen für die Benutzbarkeit getroffen werden.

In Folgenden werden zunächst die Datensammlung sowie die Fehlerklassifikation beschrieben. Anschließend wird das Konzept der semi-automatischen Evaluierung vorgestellt und es wird gezeigt, wie sich Fehler im Falle der Sprachsteuerung anhand des Klassifikationsschemas gezielt generieren lassen. Abschließend wird die Erweiterbarkeit auf andere Interaktionsmodalitäten diskutiert.

Datensammlung

Im Rahmen des EU-geförderten IST-Projektes INSPIRE (IST 2001-32746) wurden Interaktionen mit einem Smart-Home-System untersucht. Das INSPIRE-System gestattet die Bedienung verschiedener Hausgeräte (Leuchten, Rollos, TV, Videorekorder, Anrufbeantworter, etc.) über gesprochene Sprache. Es umfasst eine Signalvorverarbeitung, eine Spracherkennung, sprachverstehende Komponenten, einen Dialogmanager, eine Geräteschnittstelle sowie eine Sprachausgabe aus Bausteinen vorher aufgezeichneter Sprache. Für den beschriebenen Versuch stand ein Prototyp zur Verfügung, bei dem der Spracherkenner durch eine Online-Transkription ersetzt wurde.

24 Versuchspersonen (19-29 Jahre) führten jeweils 3 Dialoge mit dem System. Die Dialoge folgten Szenarien, die jeweils 9-11 unterschiedliche Bedienaufgaben umfassten. Nach jedem Dialog beurteilten die Versuchspersonen jeweils 37 Aussagen zu unterschiedlichen Qualitätsaspekten auf einem Fragebogen. Die Interaktionen wurden aufgezeichnet und durch einen Experten nach einem speziell zu diesem Zweck entwickelten Klassifikationsschema annotiert. Details zur Versuchsdurchführung, zu den Fragebögen und zur Annotierung finden sich in [1].

Fehlerklassifikation

Fehler werden hier als Abweichungen von einem „optimalen" Pfad durch die Interaktion zwischen Nutzer und System gesehen. Sofern ein Nutzer durch einen Dialogschritt (also eine Äußerung gegenüber dem System) von diesem „optimalen" Pfad abweicht, verlängert sich die Interaktion, d.h. der Nutzer ist seinem Ziel durch den Dialogschritt nicht näher gekommen. Diese Situation bezeichnen wir als „Fehler", obwohl dem Benutzer damit keinerlei Schuld zugewiesen werden soll. Eine solche Definition von „Fehlern" ist auch nur für aufgaben-orientierte Dialogsysteme möglich; dies ist jedoch bei den meisten kommerziell verfügbaren Systemen der Fall.

Gemäß dieser Definition wird fehlerhaftes Benutzerverhalten auf 5 Ebenen annotiert [2]:

- *Ziel-Ebene*: Dieser Fehler tritt auf, wenn das System nicht das erwartete Ausmaß an Funktionalität besitzt.
- *Aufgaben-Ebene*: In diesem Fall weiß der Benutzer nicht, wie er die Aufgabe mit Hilfe des Systems lösen kann, bspw. indem er ein Kommando zum falschen Zeitpunkt äußert, das Kommando den Dialog nicht zum Aufgaben-Ziel führt, etc.
- *Konzept-Ebene*: Der Nutzer geht von einer anderen Repräsentation der „Welt" aus, d.h. er hat ein anderes Modell vom Funktionieren des Systems in der Umgebung.
- *Kommando-Ebene*: Der Nutzer verwendet Vokabular oder grammatikalische Konstruktionen, die das System nicht erkennen oder verstehen kann.
- *Erkennungs-Ebene*: Diese Fehler sind allein der Spracherkennungskomponente zuzuschreiben, und sind auch für einen optimal erfahrenen Nutzer unvermeidlich.

Die negativen Konsequenzen der Fehler lassen sich ebenfalls kategorisieren in

- *Stagnation*, mit Sonderfall Wiederholung,
- *Rückschritt*, mit Sonderfall Neustart, sowie
- *teilweiser Fortschritt* trotz Fehlers.

Die 72 Dialoge wurden von einem Experten bzgl. der o.a. Fehlerklassen und Konsequenzen annotiert. Nach Löschung einiger fehlerhaft aufgezeichneter Daten ergaben sich insgesamt 2343 annotierte Paare Systemäußerung-Nutzeräußerung. In 28% dieser Interaktionspaare fand sich mindestens ein Fehler. Die meisten Fehler äußerten sich auf der Kommando- (51%) sowie auf der Aufgaben-Ebene (44%). Kommandofehler entstanden meist durch nicht vorgesehenes Vokabular, weniger durch dessen grammatikalische Variationen. Aufgaben-Fehler entstanden durch (1) Äußerung eines Kommandos in einem Systemzustand, in dem dieses Kommando nicht erkannt wird, (2) Verwendung eines falschen Typs (bspw. Angabe des Filmtitels anstelle der entsprechenden Listennummer), oder (3) gleichzeitige Angabe von zwei miteinander verwandten Kommandos. Bspw. versuchten Nutzer, „zwei Lampen" gleichzeitig einzuschalten; das System kann aber nur einzelne Lampen oder alle 3 Lampen gleichzeitig bedienen. Hier macht sich also eine nicht konsequent gewählte Funktionalität des Systems negativ bemerkbar.

Semi-automatische Evaluierung

Die in der Interaktion festgestellten Fehler sollen nun in einer Werkbank zu halbautomatischen Evaluierung des Dialogsystems implementiert werden, vgl. [3]. Das Prinzip dieser Werkbank ist in Abb. 1 skizziert.

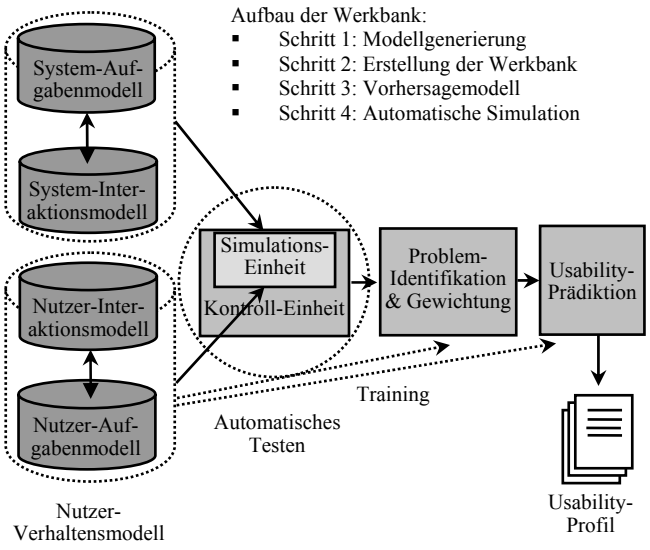

Abbildung 1: Prinzip der halbautomatischen Evaluierung mittels System- und Benutzermodellierung.

Das Verhalten des Systems wird zunächst durch ein Modell beschrieben. Dieses besteht aus einer Beschreibung der Aufgabe, die mit Hilfe des Systems gelöst werden kann, sowie des dazu implementierten Interaktionsverlaufes. Komplementär dazu wird ein Modell des Benutzerverhaltens erstellt. Dieses besteht ebenfalls aus einer Beschreibung der Aufgabe, die der Benutzer mit Hilfe des Systems lösen möchte, sowie einer Beschreibung seines Interaktionsverhaltens. Beide Modelle interagieren nun in einer automatischen Testeinheit, wobei eine Vielzahl potentieller Pfade durch den Dialog simuliert wird. Diese Pfade werden automatisch aufgezeichnet, und aus den Log-Daten wird ein Profil der zu erwartenden Qualität und Gebrauchstauglichkeit vorhergesagt.

Zur Simulation des Benutzerverhaltens wird zunächst ein „idealer" Pfad durch die Interaktion angenommen, wie er durch das Systemmodell definiert ist. Von diesem idealen Pfad werden nun regelbasiert Abweichungen generiert, die sich an den o.a. Fehlerklassen orientieren. Bspw. lassen sich Ziel-Fehler durch Verfolgung einer Aufgabe, die nicht vom System unterstützt wird, generieren. Aufgaben-Fehler lassen sich durch ansonsten interpretierbare Benutzeräußerungen an einer nicht vom System vorgesehenen Stelle im Dialogablauf generieren. Kommando-Fehler lassen sich durch Verwendung eines nicht vorgesehenen Vokabulars oder grammatikalischer Konstruktionen generieren. Erkennungsfehler lassen sich durch gezielte Vertauschung einzelner Wörter der simulierten Benutzeräußerung generieren. In der Interaktion mit dem Systemmodell treten die Konsequenzen der Fehler zutage, aus denen Rückschlüsse auf die Gewichtung der Fehler für die vom Benutzer erfahrene Qualität und Gebrauchstauglichkeit gezogen werden können.

Das Klassifikationsschema lässt sich prinzipiell auch auf andere Dialogsysteme übertragen, bspw. auf GUIs. Hier lassen sich z.B. fehlerhaftes Klicken oder falsche Feldeinträge mittels ähnlicher Regeln generieren. Zur realistischen Simulation müssen in allen Fällen Regeln gefunden werden, die einen entsprechenden Fehler triggern. Solche Wahrscheinlichkeiten werden derzeit aus Versuchen sowie aus Experteninterviews abgeleitet.

Zusammenfassung und Ausblick

Es wurde ein neues halbautomatisches Evaluierungsverfahren zur Bestimmung der Benutzbarkeit und Qualität einer intelligenten Sprachsteuerung für Hausgeräte vorgestellt. Das Verfahren beruht auf der Klassifikation und anschließender Simulation von Fehlern, wie sie in natürlichen Interaktionen festgestellt wurden. Es lässt sich auch auf andere Interaktionsmodalitäten anwenden.

Literatur

[1] Möller, S., Smeele, P., Boland, H., Krebber, J.: Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study. Computer Speech and Language 21 (2007), 26-53.

[2] Oulasvirta, A., Möller, S., Engelbrecht, K., Jameson, A.: The Relationship of User Errors to Perceived Usability of a Spoken Dialogue System. Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin (2006), 61-67.

[3] Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N.: MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations. Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh PA (2006), 1786-1789.