

Eine multimodale Suchmaschine für Nachrichtentexte: Architektur und Benutzer-Vorstudie

Alexandra Klein Ingrid Schwank Michel Génèreux Harald Trost
Österreichisches Forschungsinstitut für Artificial Intelligence (ÖFAI)¹,
Wien

Michael Trimmel

Institut für Umwelthygiene, Universität Wien

Abstract. Die Interaktion mit dem World Wide Web bietet Benutzern derzeit nur wenig echte Interaktionsmöglichkeiten, da sich Benutzereingaben meist auf Mausclicks und das Ausfüllen von Formularen beschränken. Durch das Zufügen von Sprachverarbeitungsmechanismen auf verschiedenen Ebenen kann hingegen die Kluft zwischen Navigation und Interaktion verringert werden, was zu einer Verbesserung der Kommunikationssituation führen sollte, da Benutzer auf diese Weise in freien gesprochenen und geschriebenen Äußerungen zusätzlich zu Mausclicks auf die Informationen im WWW zugreifen können. Die Repräsentation und der Austausch von Informationen aus verschiedenen Eingabemodalitäten erfordern dabei eine besonders flexible Systemarchitektur. Eine Wizard-of-Oz-Benutzervorstudie hat gezeigt, daß eine solche Erweiterung der Eingabemöglichkeiten tatsächlich zu einer höheren Akzeptanz (bei leicht verringertem zeitlichen Aufwand) führt.

Einleitung

Während Entwickler von WWW-Seiten die Möglichkeiten von Multimedia beliebig ausschöpfen und beliebige Kombinationen von Text, Ton, Bildern und Video in ihren Präsentationen verwenden können, um die Benutzer möglichst effektiv anzusprechen, sind die Reaktionen von Benutzern weitgehend auf das Ausfüllen von Formularen und Point-and-Click-Operationen beschränkt. Komplexere Interaktionsarten sind allerdings mit Mausclicks dem Ausfüllen von

¹ Das Österreichische Forschungsinstitut für Artificial Intelligence wird vom österreichischen Bundesministerium für Bildung, Wissenschaft und Kultur unterstützt. Der Fonds zur Förderung der wissenschaftlichen Forschung (FWF) fördert das beschriebene Projekt unter dem Kennzeichen P13704-INF.

Formularfeldern nicht denkbar [Cohen 1995]. Durch das Zufügen von Sprachverarbeitungsmechanismen auf verschiedenen Ebenen kann hingegen die Kluft zwischen Navigation und Interaktion verringert und somit die Kommunikationssituation verbessert werden. Freie geschriebene oder gesprochene Eingaben erlauben außerdem gezielteres Navigieren durch Hypertext-Strukturen direkt zum gesuchten Dokument. Das befreit die Suche von ihrer Abhängigkeit von der vorgegebenen Dokumentstruktur, was vorteilhaft ist, da die Intentionen von Benutzern in vielen Fällen durchaus anders sein können als die Intentionen, die bei der Erstellung von Hypertext-Strukturen von den Autoren angenommen worden sind. Weiterhin muß sich der Benutzer nicht bestimmten vordefinierten Formulierungen bei der Suche anpassen.

Multimodale Interaktion beim Zugriff auf Web-Seiten

Das vom Fonds zur Förderung der wissenschaftlichen Forschung (FWF) finanzierte Projekt "Ein multimodales Sprachinterface zum Zugriff auf Web-Seiten" untersucht Möglichkeiten der Integration von freier gesprochener und geschriebener Sprache mit klassischen Zugriffsmethoden, einschließlich der Vor- und Nachteile der verschiedenen Kombinationen. Mit Rücksicht auf die Qualität der Spracherkennung haben wir zunächst eine eingeschränkte Domäne ausgesucht, nämlich die Suche in deutschsprachigen Nachrichtentexten im WWW. Dabei verfolgt das Projekt drei Ziele. Zum einen werden empirisch fundierte Pre-design-Studien durchgeführt, die Aufschluß geben über die Rolle von freier Sprache in einem multimodalen System zum Zugriff auf Web-Seiten. Weiterhin wird im Rahmen des Projekts untersucht, wie gesprochene und geschriebene Äußerungen im Kontext eines multimodalen Systems analysiert werden können. Denn während Spracherkennung für spontansprachliche Äußerungen häufig zu hohen Fehlerraten führt, ist es möglich, durch eine Analyse des Situationskontextes der Äußerung die tatsächlich gesprochene Wortkette besser zu bestimmen. Außerdem wird im Rahmen des Projekts ein Prototypen-System entwickelt, das gesprochene und getippte Eingaben akzeptiert.

Seit einiger Zeit liegt der Schwerpunkt der Forschung auf der Integration von verschiedenen Eingabemodi, wobei gesprochener Sprache eine besondere Rolle zukommt. In solchen Systemen kann Browser-Funktionalität mit Anfragen in gesprochener Sprache gesteuert werden [Hemphill 1995, Lau 1997], doch sind solche Systeme noch weit von einer echten Integration

von Anfragen bezüglich Form, Inhalt und Struktur entfernt, wie sie in typischen Interaktionen oft vorkommen. Daher untersucht unser Projekt die benutzerfreundliche Einbindung von gesprochener und geschriebener Sprache in klassische Eingabefunktionalität. Es ist also notwendig, für das Projekt eine Benutzer-Oberfläche zu entwickeln, die Eingaben in gesprochener Sprache angemessen verarbeiten kann.

Benutzer-Vorstudie

Ein erster Schritt war der Entwurf einer Benutzeroberfläche für empirische Tests im Rahmen von Wizard-of-Oz-Studien. Wizard-of-Oz-Studien sind eine in der Entwicklung von natürlichsprachlichen Systemen bewährte Methode, um die Benutzbarkeit und Funktionalität von zukünftigen Systemen in einer Benutzer-Vorstudie zu testen. Zu diesem Zweck wird zunächst eine Benutzeroberfläche entworfen. In kontrollierten Experimenten wird Versuchspersonen erklärt, daß sie bereits mit dem realisierten System arbeiten. In Wirklichkeit werden die Interaktionen der Maschine von menschlichen "Wizards" simuliert. Anhand der Aktionen und Einschätzungen der Benutzer lassen sich Aufschlüsse darüber gewinnen, wie sich die Benutzeroberfläche gestalten bzw. verbessern läßt und wie Benutzer mit dem wirklichen System interagieren werden.

Ein Problem von Wizard-of-Oz-Studien ist die Simulation der Sprachverarbeitung. Verstehensfehler lassen sich auf der Ebene des Sprachverstehens und vor allem auf der der Spracherkennung nur sehr schlecht simulieren. Dabei bilden aber Verstehensfehler einen wichtigen Aspekt in der Interaktion der Benutzer mit dem System, da sie die Einschätzungen und die Handlungen der Benutzer sehr stark beeinflussen. Für die im Rahmen unseres Projektes durchgeführten Experimente haben wir uns entschieden, eine perfekte Spracherkennung und ein perfektes Sprachverstehen zu simulieren, obwohl wir uns der Problematik dieser Vorgehensweise bewußt waren. Sie schien uns jedoch immer noch sinnvoller als die Simulation von Fehlern, da es sehr schwierig ist, Systemverhalten adäquat vorherzusagen, vor allem, wenn das System noch nicht fertiggestellt ist.

Für die Experimente bekamen insgesamt 43 Testpersonen Instruktionen zu dem System und

jeweils 12 Aufgaben, die gelöst werden sollten. Bei den Aufgaben handelte es sich um Suchen nach kurzen Zusammenfassungen von. Die entsprechenden Artikel sollten dann von den Benutzern gesucht werden. In der Vorgehensweise der Suche funktionierte das System wie eine Suchmaschine oder ein Archiv mit Suchfunktion, nur konnten die Benutzer ihre Eingaben auch in völlig freier gesprochener oder geschriebener Sprache ausdrücken. In unterschiedlichen Abfolgen hatten alle Benutzer die Aufgabe, die vordefinierten Zeitungstexte in drei verschiedenen Kombinationen von Eingabemodi zu suchen. Eine Variante war die Verwendung von gesprochener Sprache und Mausclicks, die zweite Variante war die Verwendung von geschriebener Sprache und Mausclicks, und die dritte Variante war die Verwendung von geschriebener und gesprochener Sprache und Mausclicks nach Belieben. Bei allen Testpersonen wurde in einem Erhebungsbogen die Vorerfahrung im Bereich Umgang mit Suchmaschinen und dem Internet erhoben. Basierend auf diesen Angaben wurden die Testpersonen dann in zwei Gruppen aufgeteilt, in "Laien" und "Experten". Nach Beendigung des Experiments füllten die Testpersonen einen Fragebogen zu ihrer Einschätzung des Systems sowie zu den Kombinationen von Eingabemöglichkeiten aus. So ist es möglich, die Beurteilungen von den Testpersonen mit den Zeiten zu vergleichen, die die Personen unter den verschiedenen Rahmenbedingungen für die Bearbeitung der Aufgaben benötigten.

Insgesamt zeigten sich die Benutzer zufrieden mit der Interaktion mit dem System; auf einer Werteskala von 1 (beste Bewertung) bis 5 (schlechteste) wurde im Mittelwert die 2,049 vergeben. Dabei waren Laien im allgemeiner zufriedener mit dem System; sie vergaben im Mittel die Note 1,8, die Experten die Note 2,3. Dies stützt die These, daß die Kombination von traditionellen Eingabemodalitäten mit freier geschriebener und getippter Sprache besonders bei Nicht-Experten auf große Akzeptanz stößt, weil dies eine natürlichere Form der Kommunikation darstellt. Es stellte sich weiterhin heraus, daß die Testpersonen bei der Variante mit gesprochener und geschriebener Sprache bzw. nur mit gesprochener Sprache im Schnitt weniger Zeit für die Bearbeitung der Aufgaben benötigten als für die Variante nur mit geschriebener Sprache. Die Auswertungen sind noch nicht abgeschlossen, doch deuten die ersten Ergebnisse darauf hin, daß multimodale Interaktion besonders für Laien einen benutzerfreundlichen und effizienten Zugang zu Informationen im WWW darstellt. Dies ist besonders wichtig, da Informationen aus

dem Internet für alle zu einer Quelle von Informationen im privaten und beruflichen Bereich geworden sind, zu der der Zugang möglichst unkompliziert sein sollte, nicht nur aus technischer, sondern auch aus kommunikativer Sicht.

Ausblick

Aus der Perspektive der Sprachverarbeitung ergeben sich Herausforderungen im Bereich der Architektur, der Wissensrepräsentation und der Verwaltung von Informationen, da die aus Benutzereingaben extrahierten Informationen jeweils mit der Kommunikationssituation abgeglichen werden müssen. Der durch die Tests mit Benutzern entstandene Korpus bildet dabei eine gute Grundlage für die Simulation von echten komplexen Benutzerinteraktionen. Damit die Benutzer-Eingaben richtig interpretiert werden können, müssen die Einschränkungen, die durch die kommunikative Situation gegeben sind, identifiziert und repräsentiert werden. Dies spricht für ein aktionszentriertes Modell, welches Benutzerhandlungen in Form von geschriebener oder gesprochener Sprache oder Mausclicks als Instanzen von Informationsanfragen in spezifischen kommunikativen Umgebungen behandelt [Johnston 1998]. Auf diese Weise sind Benutzerhandlungen weit weniger eingeschränkt als die Mausclicks in graphischen Benutzeroberflächen, und die Interaktion mit Informationen auf Web-Seiten gewinnt eine neue kommunikative Dimension, die besonders für Laien, aber auch für geübte Benutzer in höherer Akzeptanz resultiert, wie unsere Benutzer-Vorstudie gezeigt hat.

Literatur

- P.H. Cohen, S.L. Oviatt, The Role of Voice Input for Human-Machine Communication. Proceedings of the National Academy of Sciences, 92(22):9921-9927, 1995.
- C.T. Hemphill, P.R. Thrift, Surfing the Web by Voice. In Proc. ACM Multimedia, pp. 215-222, San Francisco, CA, November 1995.
- M. Johnston, Unification-based multimodal parsing. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 624-630, Université de Montréal, Quebec, Canada, August 10-14, 1998.
- R. Lau, G. Flammia, C. Pao, V. Zue, Web-GALAXY – Integrating Spoken Language and Hypertext Navigation. In Proceedings of EUROSPEECH 1997, pp. 883-996, Rhodos, Griechenland, September 1997.