DOMÄNENSPEZIFISCHE OPTIMIERUNG EINES BIDIREKTIONALEN SPRACHÜBERSETZUNGSSYSTEMS IN VERTEILTER ARCHITEKTUR

Michael Stier und Stefan Feldes

Institut für Digitale Signalverarbeitung
Fachhochschule Mannheim - Hochschule für Technik und Gestaltung
mstier@gmx.de, s.feldes@fh-mannheim.de

Abstract: Im Rahmen dieser Arbeit wird untersucht, wie die Qualität eines Sprachübersetzungssystems durch gezielte Optimierung auf eine Domäne gesteigert werden kann. Durchgeführt wurden die Untersuchungen an einem Experimentalsystem, das bewusst nur auf der Basis käuflicher Produkte zur Spracherkennung, Übersetzung und Sprachsynthese aufgebaut wurde, wobei die verteilte Architektur des Systems das Internet als Verbindung nutzt. An diesem System wurden die grundsätzliche Machbarkeit und das notwendige Vorgehen bei der Ausrichtung und Optimierung auf die gewählte Domäne der Hotelreservierung untersucht. Durchgeführt wurde die Optimierung zunächst für die Spracherkennung durch sprecherabhängige Anpassung des akustischen Modells, domänenspezifische Anpassung des Sprachmodells zum einen in Form eines statistischen Modells, sowie zum anderen in Form einer kontextfreien Grammatik. Die Evaluierung durch Konversationstests zeigte eine deutliche Verbesserung. Beispielsweise wurde bei Nutzung des statistischen Sprachmodells die Wortakkuratheit von 75% auf 92% gesteigert.

1 Einleitung

Die automatische Erkennung und Übersetzung von Spontansprache, wie sie in einem freien Dialog üblich ist, gilt als eine der anspruchvollsten Forschungsaufgaben innerhalb der Sprachtechnologie. Die Aufgabe umfasst, dass kontinuierlich gesprochene Sprache in beiden verwendeten Sprachen erkannt, in die jeweils andere Sprache übersetzt und synthetisch ausgegeben wird. Somit wird eine Möglichkeit der Verständigung eröffnet, wo kein Übersetzer verfügbar ist. Intensive Forschung wurde in den letzten Jahren auf diesem Gebiet betrieben. Groß angelegte Projekte, wie beispielsweise Verbmobil, haben wichtige wissenschaftliche Fortschritte gebracht. Mit einer sehr komplexen Systemarchitektur und aufwendigen statistischen sowie linguistischen Verfahren zur Analyse und Übersetzung konnte mit Verbmobil [1] eine Übersetzungsleistung von approximativ 80% richtigen Übersetzungen erreicht werden.

Bedingt durch die stetig fortschreitende Entwicklung erreichen inzwischen auch handelsübliche kommerzielle Produkte zur Spracherkennung sowie Systeme zur textuellen Übersetzung brauchbare Ergebnisse. Die Idee liegt nahe, ein System zur Sprachübersetzung zu entwerfen, welches auf einer geeigneten Kombination derartiger käuflicher Komponenten basiert. Wichtig ist jedoch beim derzeitigen, begrenzten Leistungsstand der Komponenten ein darauf abgestimmtes Anwendungsszenario sowie eine anpassungsfähige Systemarchitektur zu entwickeln. In einer früheren Arbeit [2] wurde das Konzept eines über Telefon erreichbaren *zentralen Übersetzungssystems* verfolgt. Dieses konnte per Telefonkonferenzschaltung von den beiden verschiedensprachigen Teilnehmern als Dolmetscheinheit angewählt werden. Andere Ansätze hingegen (z.B. [3]) streben an, das komplette System auf einem mobilen *Endgerät* (PDA) zu realisieren, was natürlich die Schwierigkeit der begrenzten Ressourcen auf derartigen Geräten mit sich bringt. Die Suche nach einer geeigneten und einfach zu realisierenden Architektur führte uns auf ein *verteiltes System*, in das kommerzielle Produkte gut integriert werden können sowie die Infrastruktur des Internets vorteilhaft genutzt werden kann. Dieses Konzept [4] wurde in Form eines Experimentalsystems realisiert. Es arbeitet

bidirektional, wobei jeder der beiden Gesprächspartner seinen Multimedia-PC mit lokal verfügbarer Spracherkennung und Sprachsynthese nutzt. Über eine IP-Anbindung wird die jeweils erkannte Äußerung in Textform an einen Server zur Übersetzung geschickt, der wiederum den übersetzten Text an den zweiten Gesprächspartner zur dortigen Synthese weiterleitet. Dieses System dient als Grundlage für die im Folgenden beschriebenen Untersuchungen, die sich insbesondere darauf konzentrieren, wie die Qualität eines solchen Sprachübersetzungssystems durch gezielte Optimierung auf eine Anwendungsdomäne gesteigert werden kann.

Im folgenden Abschnitt beginnen wir mit der Analyse der Anwendungsszenarien. Die sich daraus ergebende Domäne stellt die Ausgangslage für das in Abschnitt 3 beschriebene System dar. Anschließend wird die Vorgehensweise bei der Optimierung und die durchgeführten Experimente beschrieben. Am Ende des Beitrags steht die Diskussion der Ergebnisse.

2 Anwendungsszenario

Die anspruchsvolle Aufgabe der bidirektionalen Sprachübersetzung von Spontansprache kann nach wie vor nur gelöst werden, wenn der Dialog auf ein vorgegebenes thematisches Umfeld (Domäne) beschränkt wird. Für den Erfolg eines Sprachübersetzungssystems ist dann neben der technischen Leistungsfähigkeit des Systems insbesondere die Attraktivität des Anwendungsszenarios ausschlaggebend. Wesentlich sind dabei die unmittelbar einsichtige Nützlichkeit des Übersetzungsdienstes im jeweiligen Anwendungskontext, die intuitiv erfassbare Funktionalität des Dienstes sowie eine einfache Bedienbarkeit.

Bei unserer Betrachtung möglicher Anwendungsszenarien für die Übersetzung haben wir bewusst auf den Fall der Dialogführung zwischen Mensch und Maschine verzichtet, da in diesem Fall die Notwendigkeit einer automatischen Übersetzung in der Regel nicht gegeben ist, sondern direkt durch ein multilingual ausgelegtes Dialogsystem realisiert werden kann.

Ein von uns betrachtetes Übersetzungssystem dagegen übernimmt die Aufgabe eines Dolmetschers für einen Dialog von Mensch zu Mensch gemäß Abbildung 1. Es übersetzt die spontansprachlichen Äußerungen des Teilnehmers 1 (Kunde) von seiner Muttersprache (im Beispiel deutsch) in eine Fremdsprache (englisch). Der Gesprächspartner 2 hört die übersetzte Äußerung und antwortet seinerseits in der Muttersprache. Diese Äußerung wird wiederum übersetzt. Dabei greift das automatische Übersetzungssystem nicht in das Gespräch zwischen den beiden Gesprächspartnern ein. Sinnvolle Anwendungsszenarien ergeben sich also überall dort, wo der persönliche Austausch der Gesprächspartner unterschiedlicher Muttersprache wesentlich ist.

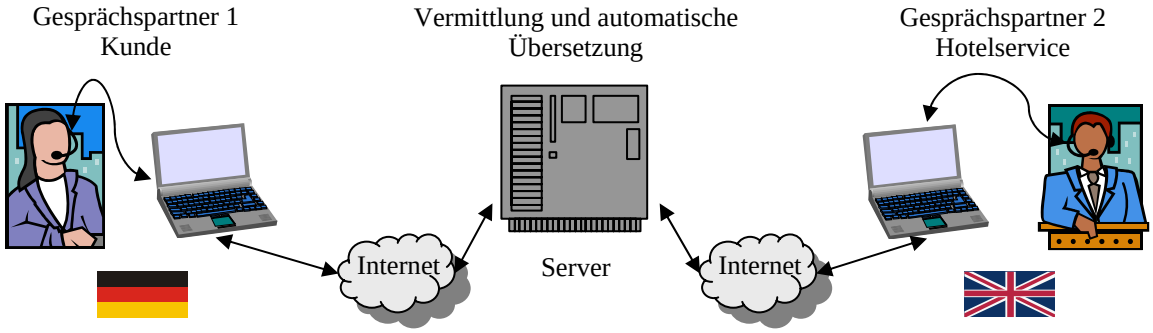

Abbildung 1 - Automatische Übersetzung in verteilter Architektur

Genauer betrachtet spielt für die Wahl eines geeigneten Anwendungsszenarios auch die Systemarchitektur eine wichtige Rolle. Die drei in der Einleitung genannten Ansätze sollen diesbezüglich im Folgenden kurz verglichen werden.

Nützliche Anwendungen für einen *zentralen telefonbasierten Übersetzungsdienst* wie in [2] beschrieben profitieren vor allem von der unkomplizierten Verfügbarkeit des Telefons als Endgerät, auch unterwegs. Das Szenario der Taxibestellung oder der Notarztruf bei einer plötzlichen Erkrankung im Ausland sind dafür Beispiele.

Die Implementierung des Übersetzungssystems auf einem mobilen *Endgerät* (z.B. PDA), ist insbesondere dann sinnvoll, wenn die Gesprächspartner am gleichen Ort beieinander sind und somit über dieses Gerät kommunizieren können. Ein Beispiel hierfür sind medizinische Interviews von Ärzten bei der Notfallversorgung in Krisengebieten [3].

Der hier untersuchte Ansatz mit *verteilter Systemarchitektur* ist hingegen passend für Szenarien, wenn die Gesprächspartner am Arbeitsplatz sind oder im häuslichen Bereich am PC sitzen. Ein vorteilhaftes Szenario kann sich hier insbesondere im Zusammenspiel mit der Nutzung des World Wide Webs ergeben. Die Wahl fiel daher auf den Bereich der Reiseplanung, die neben der Nutzung des Webs auch verschiedentlich Bedarf für Sprachkommunikation mit Übersetzung zeigt. Deshalb haben wir das Anwendungsszenario *Hotelreservierung* aus dem Bereich der *Reiseplanung* ausgewählt.

Nach der Recherche von geeigneten Hotels oder gar nach der Buchung eines Hotels über das Web für eine Geschäftsreise in das Ausland kann aus vielerlei Gründen die Notwendigkeit entstehen, in einem persönlichen Gespräch mit dem Hotel Rücksprache zu halten, beispielsweise bei einer Umbuchung oder zur Stornierung der Reservierung.

Im Beispiel einer Umbuchung könnte sich folgendes Szenario abspielen:

Gebucht sind drei Doppelzimmer. Krankheitsbedingt kann ein Teilnehmer nicht mitreisen. Somit kommt die Frage auf, ob es sinnvoll ist eines der Doppelzimmer umzubuchen in ein Einzelzimmer oder entstehen durch die Umbuchung Gebühren in einer Höhe, die in diesem Falle nicht vertretbar wären.

Derartige Fragen können in einem persönlichen Beratungsgespräch durch das Hotel individuell geklärt werden. Dies ist nur ein Beispiel, das die Nützlichkeit des Übersetzungssystems aufzeigen soll. Neben weiteren Szenarien findet das hier geschilderte Beispiel der Umbuchung Verwendung als Aufgabenstellung bei den durchgeführten Experimenten zur Spracherkennung.

3 System-Architektur

Im Rahmen dieser Arbeit wurde das in [4] konzipierte Experimentalsystem weiterentwickelt [5], welches als Basis für die Untersuchungen verwendet wurde. Die Gesamtarchitektur wird in Abbildung 2 grafisch dargestellt. Es handelt sich um ein System mit modularer und verteilter Architektur. Der jeweilige Gesprächspartner nutzt seinen Multimedia-PC mit Headset, der innerhalb der verteilten Architektur als Client fungiert. Dieser benutzerspezifische Client enthält Module zur Spracherkennung und Sprachsynthese für die jeweilige Muttersprache des Anwenders. Die Verwendung des PCs erlaubt es gemäß dem Paradigma „Personal Computer“ diesen auch auf die Stimme des Benutzers zu konfigurieren, wie dies bei einem benutzerspezifischen PC in vielerlei Hinsicht praktiziert wird. Somit wird die Spracherkennung sprecherabhängig, wodurch prinzipiell eine höhere Wortakkuratheit erreicht werden kann. Als drittes Modul enthält der Client ein Verbindungsmodul zur Kommunikation mit dem Server über eine IP-Anbindung. Alle Clientmodule sind eigenständige Komponenten und kommunizieren untereinander über Socketverbindungen, wobei die Steuerung durch das Verbindungsmodul übernommen wird. Der Client baut eine Verbindung zu einem Server auf, der die Vermittlung zum Gesprächspartner und dessen Client übernimmt. Die Übersetzung der Äußerung für das jeweilige Sprachpaar wird von einem Übersetzungsserver übernommen, der durch den Vermittlungsserver kontaktiert wird. Hiermit ist die Möglichkeit der einfachen

Erweiterung um weitere Sprachpaare zur Übersetzung hin zum multilingualen System gegeben.

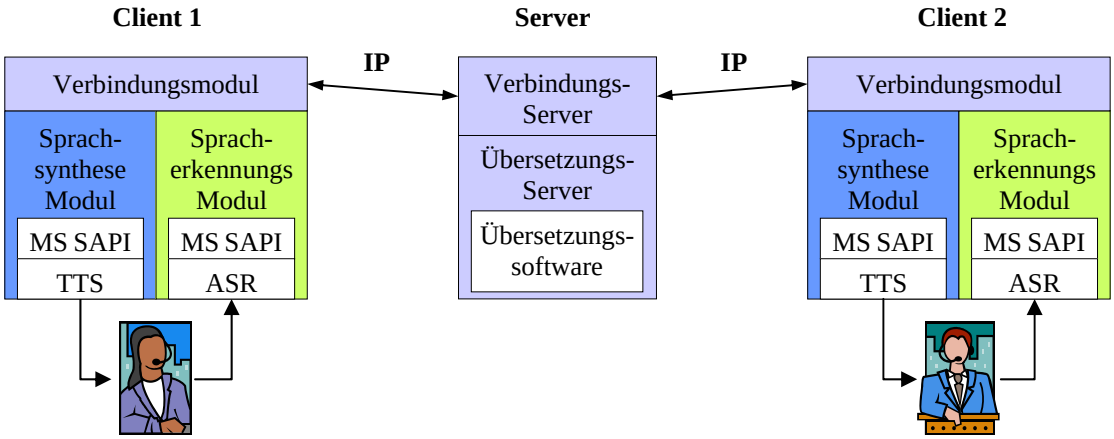

Abbildung 2 - Schematische Darstellung der prinzipiellen Systemarchitektur

Im Folgenden wird ein kurzer Überblick über die Clientmodule gegeben:

- Das **Spracherkennungsmodul** beinhaltet sämtliche Funktionalität zur Spracherkennung. Es besteht die Möglichkeit der Auswahl eines lokal installierten Erkenners (ASR). Für diesen Erkenner kann das benutzerspezifische Profil des Anwenders verwendet werden, wobei hier noch zwischen statistischem Sprachmodell oder einer kontextfreien Grammatik gewählt werden kann. Zur Auswertung des Erkennungsergebnisses kann das Audiologging und Protokollieren der Hypothese aktiviert werden. Nach dem Gesprächsaufbau kann der Anwender zwischen zwei Betriebsmodi wählen. Im „hands-free"-Modus wird die vom Erkenner gelieferte Hypothese automatisch an den Server weitergeleitet, während im „push-to-send"-Modus vor der manuellen Weiterleitung die Möglichkeit besteht, die Hypothese in einem Editierfenster zu begutachten. Das Modul wurde auf die Microsoft Speech API [6] programmiert. Somit können handelsübliche Produkte verschiedener Hersteller, die diese Schnittstelle unterstützen, eingebunden werden. Am Experimentalsystem wurden die Untersuchungen mit dem Dragon NaturallySpeaking Erkenner durchgeführt. Mit dem IBM ViaVoice Produkt wurde die Kompatibilität bestätigt. Die hier verwendeten Erkenner ermöglichen die Spracherkennung zum einen auf Basis von kontextfreien Grammatiken (CFG) zum anderen auf Basis von statistischen Sprachmodellen. Dies stellt eine Erweiterung gegenüber den Untersuchungen in [2] dar, welches lediglich die Verwendung von CFG erlaubte.

- Das Übersetzungsergebnis wird mit Hilfe des **Sprachsynthesemoduls** ausgegeben. Die zu verwendende TTS-Engine (Text-to-Speech) kann aus den lokal installierten ausgewählt werden. Die Ausgabe erfolgt über die Soundkarte an das Headset. Wieder wurde auf die MS SAPI programmiert, so dass alle kompatiblen TTS-Systeme eingesetzt werden können. Zum Einsatz kam die Realspeak TTS von Scansoft.

- Die clientseitige Steuerung der Anwendung erfolgt durch das **Verbindungsmodul**. Die übersichtlich gegliederte Oberfläche erlaubt es dem Benutzer die IP-Adresse des gewünschten Gesprächspartner anzugeben, die beiden anderen Clientmodule einzubinden, ein Gespräch aufzunehmen, anzunehmen, abzulehnen und zu beenden. Ausschließlich das Verbindungsmodul kommuniziert mit dem zentralen Vermittlungsserver.

Der zentrale Server enthält folgende Module:

- Die Kommunikation zweier Gesprächspartner findet über den **Vermittlungsserver** statt. Dieser leitet die eingehende Äußerung des einen Gesprächspartners in Textform weiter an den Übersetzungsserver und anschließend wird der übersetzte Text an den Client des anderen Gesprächspartners übertragen. Auch der Verbindungsauf- und -abbau wird durch den Vermittlungsserver koordiniert.
- Für die textuelle Übersetzung der Ausgabe des Spracherkenners wurde der L&H PowerTranslator für die Sprachpaare deutsch/englisch und englisch/deutsch verwendet. Für jedes Sprachpaar wird ein eigener **Übersetzungsserver** eingesetzt, an den der Request vom Vermittlungsserver weitergeleitet wird. Der PowerTranslator arbeitet nach dem Prinzip der direkten Übersetzung basierend auf einem umfangreichen Wörterbuch und Regelwerk. Darüber hinaus können domänenspezifische Wörterbücher angelegt werden, wodurch eine bessere Übersetzung erreicht werden kann.

4 Domänenspezifische Optimierung

Für das beschriebene System wurde eine domänenspezifische Optimierung vorgenommen. Dazu sind Anpassungen im Bereich der Spracherkennung und Übersetzung notwendig. Die Optimierung der direkten Übersetzung erfolgt durch Einpflegen unbekannter Wörter und Phrasen in das Wörterbuch. Mit der gezielten Verwendung von Attributen kann auch Einfluss auf das Regelwerk genommen werden. Zunächst konzentrierten sich die Optimierungsmaßnahmen jedoch auf den Bereich der Spracherkennung.

Bedingt durch die Verwendung von handelsüblichen Komponenten sind die Möglichkeiten der Optimierung beschränkt. Im Bereich der Spracherkennung bietet die Diktiersoftware üblicherweise an zwei Schnittstellen eine gewisse Einflussnahme an. Dies ist einmal die Anpassung des akustischen Modells und zum Zweiten die Anpassung des Sprachmodells.

Die Anpassung des **akustischen Modells** (AM) erfolgt unweigerlich mit dem Anlegen eines neuen Anwenders der Erkennungssoftware. Dazu diktiert der Anwender einen vorgegebenen Trainingstext. Die Software analysiert die Aufnahme und passt das AM selbständig an. Dieses mindest vorgeschriebene Training kann durch weitere Trainingphasen ergänzt werden, um so eine spezifischere Anpassung des Modells zu erreichen. Pflegt der Anwender neue Wörter in das Lexikon ein, ist es auch sinnvoll für diese das akustische Training durchzuführen.

Einfluss auf das **statistische Sprachmodell** (LM) kann durch die Analyse von Textdokumenten genommen werden. Dazu muss eine möglichst umfangreiche Sammlung von domänenspezifischen Äußerungen bereitgestellt werden. Bei deren Analyse werden außerdem unbekannte Wörter erkannt und dem Lexikon hinzugefügt.

Alternativ zur Verwendung eines LMs unterstützen die verwendeten Erkenner auch **kontextfreie Grammatiken**. Diese wurden durch Analyse einer Sammlung domänenspezifischer Äußerungen erstellt.

5 Konversationstests

Mit dem oben genannten Sprachübersetzungssystem wurde eine Vielzahl von Experimenten auf der Basis von Konversationen der Versuchspersonen durchgeführt. Aus der Reihe der möglichen Variationen wurde als Testszenario „deutschsprachiger Kunde und englischer Hotelservice“ ausgewählt. Die erste Phase der Optimierung konzentrierte sich auf den Bereich der Spracherkennung. Deshalb wurde folgender Experimentaufbau gewählt, der die gezielte Analyse der ASR-Effekte erlaubt und andere Einflüsse ausschließt. Die Testperson mit der Rolle des Hotelservice wurde ersetzt durch den Betreuer der Experimente. Die Kommunika-

tion mit dem System erfolgte für diesen nicht über die Sprache, sondern per Tastatur und Bildschirm. Der modulare und verteilte Aufbau des Systems erlaubt dies in einfacher Art und Weise. Mit diesem Vorgehen konnte das mögliche Fehlverhalten der Spracherkennung und Übersetzung auf der Seite des zweiten Gesprächpartners ausgeschlossen werden und der Test konzentrierte sich somit auf die Erkennung und Übersetzung des ersten Gesprächpartners, welcher von der monolingualen Durchführung des Experiments jedoch nichts bemerkt.

Für die zweite Phase sind erweiterte Tests geplant, welche mit echter bidirektionaler Konversation erfolgen sollen. Messgrößen wie die Rate der richtigen Übersetzungen, die Dialogerfolgsrate sowie die subjektive Bewertung des Systems durch die Testteilnehmer werden dann zur Beurteilung herangezogen.

Bei den hier durchgeführten Experimenten wurde die Qualität mit Hilfe der Wortakkuratheit (WA) bestimmt. Dazu wurden zwei unterschiedliche Testreihen durchgeführt. In der Testreihe A wurden vorgegebene Dialoge von den Sprechern gesprochen. Testreihe B wurde mit spontanen Dialogen aufgenommen. Dazu wurde lediglich eine bestimmte Aufgabenstellung vorgegeben, die es galt zu lösen. Ein Testdurchlauf umfasste jeweils die Dialogszenarien *Buchung*, *Stornierung*, *Bestätigung* und *Umbuchung* einer Hotelreservierung. Die Testpersonen nahmen in den Testreihen A und B jeweils die Rolle des Kunden ein. Folgende Abbildung zeigt eine Übersicht zu den durchgeführten Experimenten:

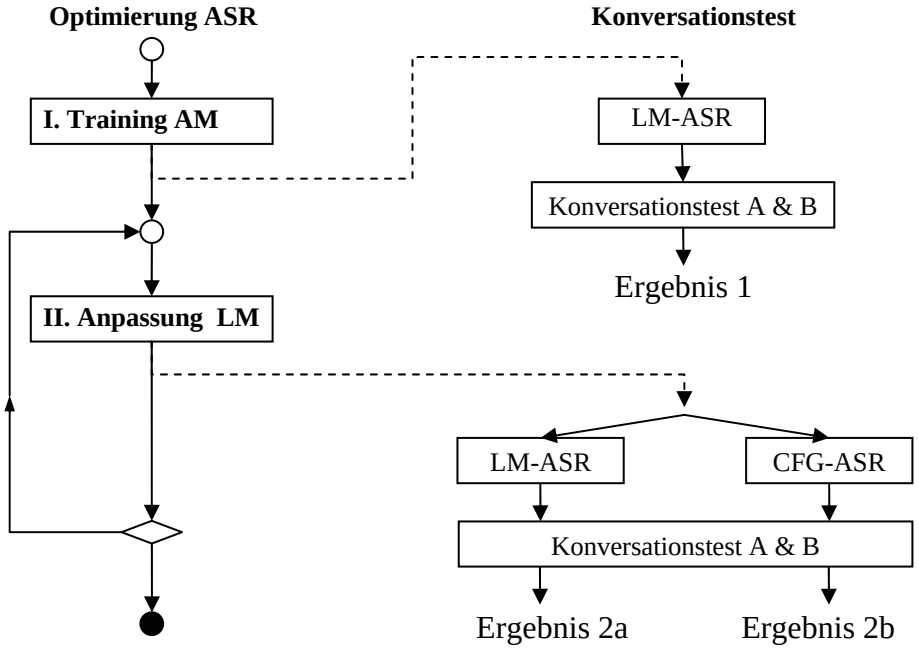

Abbildung 3 – Übersicht zu den durchgeführten Experimenten

Die Übersicht zeigt, zu welchen Schritten der Optimierung der Spracherkennung die Experimente durchgeführt wurden. Mit dieser Versuchanordnung kann die Verbesserung der Erkennungsrate mit jedem Optimierungsschritt bestimmt werden.

Wegen der Vielzahl der vorgesehenen Experimente sind mit Hilfe der Audiologging-Funktion Sprachdaten von jeder Testperson im ersten Testdurchlauf gesammelt worden. Anhand dieser Daten konnte zu jedem Optimierungsschritt ein Erkennungsdurchlauf durchgeführt werden. Bedingt durch die Verwendung desselben Sprachdatenmaterials zu jedem Experiment sind die Erkennungsergebnisse direkt vergleichbar.

Die durchgeführten Sprachaufnahmen dienten neben der Sammlung der Audiodaten auch zur Identifizierung des domänenspezifischen Vokabulars, welches zur Anpassung des Sprachmodells sowie für die Erstellung der kontextfreien Grammatik herangezogen werden konnte.

6 Ergebnisse

Zur Bewertung der Spracherkennung wurde das objektive Maß der Wortakkuratheit (WA) herangezogen. Diese wurde mit Hilfe der Levenshtein Distanz [7] auf Wortebene bestimmt.

Folgende Tabelle zeigt die Erkennungsergebnisse, die sich durch die Auswertung der Konversationstests ergeben haben. Angegeben wurden die Mittelwerte $\bar{x}_{\text{ges}}$ über alle Äußerungen einer Testreihe sowie jeweils das Konfidenzintervall $\bar{x}_{\text{t}} \pm k$ zum 90% Niveau nach der t-Statistik basierend auf den Mittelwerten der Sprecher.

WA	**nach AM Anpassung**			**nach LM Anpassung**		
für	$\bar{x}_{\text{ges}}$	$\bar{x}_{\text{t}} - k$	$\bar{x}_{\text{t}} + k$	$\bar{x}_{\text{ges}}$	$\bar{x}_{\text{t}} - k$	$\bar{x}_{\text{t}} + k$
gelesene Dialoge	74,9%	68,8%	81,0%	91,9%	88,5%	95,4%
spontane Dialoge	69,2%	61,2%	74,4%	85,2%	80,0%	89,1%

Tabelle 1 – Ergebnisse der Konversationstests bei Verwendung des statistischen Sprachmodells

Diesen Ergebnissen liegt für Testreihe A insgesamt eine Sammlung von 44 Dialogen mit 307 Äußerungen zu Grunde und für Testreihe B wurden 42 aufgabenorientierte Dialoge geführt mit insgesamt 382 Äußerungen.

7 Diskussion

Die Ergebnisse zeigen den Erfolg der Optimierung. Mit dem angepassten statistischen Sprachmodell konnte für vorgegebene Dialoge eine mittlere Verbesserung der Wortakkuratheit um 23% erreicht werden, was einer Reduktion der Fehlerrate um 68% entspricht. Bemerkenswert ist, dass auch für spontane Dialoge die gleiche relative Verbesserung erreicht werden konnte. Positiv zu vermerken ist außerdem, dass die Streuung der Ergebnisse zwischen verschiedenen Sprechern (s. Konfidenzintervall) verringert ist.

Eine erste Grammatik wurde aus der Analyse der gesammelten Dialoge entwickelt. Auf dieser Basis ergab sich für die vorgegebenen Dialoge (Testreihe A) eine mittlere Wortakkuratheit von 94% und für die spontanen Dialoge (Testreihe B) eine mittlere Wortakkuratheit von 83%. Damit ist das Ergebnis für Testreihe A wie erwartet im Vergleich zum statistischen Sprachmodell besser. Für Testreihe B trifft dies allerdings nicht zu. Eine Analyse zeigte, dass hier das Phänomen der Häsitationen Probleme aufwarf. Kurze Sprechpausen innerhalb der Äußerungen führten dazu, dass der Erkenner jeweils Teiläußerungen zwischen den Pausen detektierte und diese zurückwies, obwohl die Äußerung in zusammenhängender Form in der Grammatik berücksichtigt war. Eine erste einfache Gegenmaßnahme besteht darin, die entsprechenden Timeout-Einstellungen des Erkenners zu optimieren.

Teil der Optimierung war die Ergänzung des generischen Erkennervokabulars um domänenspezifische Worte, um „Out-Of-Vocabulary“-Effekte zu vermeiden. Dies ist bekanntermaßen bei fremdsprachigen Ausdrücken problematisch [2]. In unserem Fall waren dies insbesondere englische Orts- und Hotelnamen. Die eingesetzten Diktiererkenner erlauben jedoch, für unbekannte Worte ein akustisches Training durchzuführen. Die Auswertung der Erkennungsergebnisse zeigt, dass dies offensichtlich auch für fremdsprachige Worte gut funktioniert. Hotelnamen wurden nach Anpassung der Modelle mit sehr hoher Wahrscheinlichkeit richtig erkannt.

Für die bereits in [2] beschriebene Problematik der Syntaxinterpretation mit Hilfe von Satzzeichen konnten hier Lösungswege untersucht werden. Für die korrekte Übersetzung einer Äußerung ist es von entscheidender Bedeutung zu wissen, ob es sich um eine Frage oder

einen Aussagesatz handelt. Die Übersetzungskomponente erkennt einen Fragesatz jedoch ausschließlich am abschließenden Fragezeichen. Die Spracherkennung dagegen liefert zunächst keinerlei Satzzeichen. Bei der Verwendung kontextfreier Grammatiken kann hier relativ komfortabel Abhilfe geschaffen werden, indem die Satzzeichen in einer Annotation zugefügt werden. Im Falle des statistischen Sprachmodells steht diese Lösung leider nicht offen. Im Rahmen unseres Systems konnten aber drei Lösungsansätze auf der Ebene des User-Interfaces realisiert werden. Die erste Lösung basiert darauf, dass der Anwender die Erkennungshypothese in einem Editierfenster (s. Kap. 3) sieht und somit das Satzzeichen per Hand ergänzen kann. Die zweite Möglichkeit sieht vor, dass der Anwender die Satzzeichen - wie bei einer Diktiersoftware üblich - diktiert. Die dritte Möglichkeit besteht darin, dass der Anwender den Fragesatz mit dem Wort *Frage* beginnt. Eine Äußerung wie z.B. „*Frage haben Sie Zimmer frei*“ kann dann im Postprocessing mit einem Fragezeichen ergänzt werden. Den Testpersonen standen die drei Möglichkeiten zur Wahl. Eine einheitliche Präferenz konnte nicht festgestellt werden, aber jeder hatte sich in kurzer Zeit an eine persönliche Vorzugsart gewöhnt.

Das gleiche gilt für die alternativen Betriebsmodi „push-to-send“ und „hands-free“. Ein Teil der Anwender verwendete den „hands-free“-Modus, um die Vorteile der Spracheingabe zu nutzen und sich ganz auf den Dialog mit dem Gesprächspartner konzentrieren zu können. Der andere Teil verzichtete ganz bewusst auf die automatische Weiterleitung der Hypothese, um diese vor der Weiterverarbeitung einsehen und gegebenenfalls editieren zu können.

Abschließend kann festgestellt werden, dass die gezielte Ausrichtung des Erkenners auf die gewählte Domäne eine deutlich verbesserte Erkennungsgenauigkeit gebracht hat. Dies gilt sowohl für das statistische Sprachmodell als auch für die (hier nur ansatzweise ausgebaute) Grammatik. Zwischen beiden Ansätzen sind im Wesentlichen die folgenden Vor- und Nachteile abzuwägen: Durch die stärkere Einschränkung des Suchraums bietet der CFG-Ansatz prinzipiell bessere Chancen auf höhere Erkennungsgenauigkeit. Das statistische Sprachmodell deckt dagegen leichter eine größere Formulierungsvielfalt ab. Um den Vergleich für unseren Anwendungsfall fundierter führen zu können, soll zunächst eine möglichst umfassende Grammatik erstellt und dann vergleichende Tests im Betrieb des Experimentalsystems durchgeführt werden.

Eine hohe Erkennungssicherheit stellt eine wesentliche Grundvoraussetzung für die Qualität des Sprachübersetzungssystem insgesamt dar. Ebenso wichtig ist jedoch die Domänenoptimierung der Übersetzungskomponente, die als weiterführende Aufgabe zu lösen ist.

Diese Arbeiten wurden unterstützt durch das Förderprogramm 2004 der Karl-Völker-Stiftung, Mannheim

Literatur

[1] Wahlster, W.: “Verbmobil: Foundations of speech-to-speech Translation”, Springer Verlag, 2000.
[2] Thielemann, C.; Feldes, S.; Luber, C.: „Entwicklung und Einsatz automatischer Sprach-Übersetzung für Telefon-Dienste“. In ESSV 2001, Cottbus, S. 159-166
[3] Waibel, A. et al.: „Speechalator: two-way speech-to-speech translation on a consumer PDA”. In Eurospeech 2003, Geneva, S. 369-372
[4] Fogang, E.: „Entwurf und Implementierung einer verteilten Architektur zur automatischen Übersetzung gesprochener Sprache“, Diplomarbeit an der FH-Mannheim, 2002.
[5] Stier, M.: “Domänenspezifische Optimierung eines bidirektionalen Sprachübersetzungssystems in verteilter Architektur”, Masterarbeit an der FH-Mannheim, 2004.
[6] Microsoft Speech API: http://www.microsoft.com/speech/
[7] Okuda, T.; Tanaka, E. und Kasai, T. (1976): „A Method fort he Correction of Garbled Words based on the Levinshtein Metric“. In: IEEE Transactions, Bd. 25(2), S. 172-178