Abschlussbericht

COLLATE-UdS

Computational Linguistics and Language Technology for Real-World Applications

Universität des Saarlandes
FR 4.7 Computerlinguistik
und
FR 6.7 Informatik
66041 Saarbrücken

Prof. Dr. Manfred Pinkal
Prof. Dr. Hans Uszkoreit
Prof. Dr. Wolfgang Wahlster
Dr. Gregor Erbach
Christian Braun
Gerhard Fliedner
Christian Müller
Dr. C. J. Rupp
Rainer Wasinger
Tianfang Yao
Zhiping Zheng

Juni 2004

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IN A01 B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

I Kurzdarstellung des Projekts

I.1 Aufgabenstellung

Gegenstand des Vorhabens ist die Vorlauf- und Anschubforschung eines nationalen Kompetenzzentrums für Sprachtechnologien, die noch nicht am Markt vorhanden sind. Die Kompetenz soll der F&E-Gemeinschaft, den deutschen Firmen, die sprachtechnologische Produkte entwickeln und vermarkten, sowie potenziellen Anwendern zugute kommen. Sie soll aber auch Saarbrücken und damit Deutschland im internationalen Wettbewerb als herausragenden F&E-Standort der Sprachtechnologie stärken.

Das Projekt COLLATE dient der Stärkung der internationalen Stellung der deutschen F&E in dem strategisch wichtigen Bereich der Sprachtechnologie. Sprachtechnologie ist eine wesentliche Basiskomponente für zukünftige Informations- und Kommunikationstechnologien. Ein besonderer Schwerpunkt des Projekts ist der Brückenschlag von aktuellen Sprachtechnologieforschung hin zur industriellen Verwertung. Das Hauptziel ist dabei die Beschleunigung des Transfers aus der Forschung in praktische Anwendungen. Zur Erreichung dieses umfassenden und anspruchsvollen Ziels besteht das Projekt aus mehreren komplementären, interagierenden Komponenten.

Ein zentraler Bestandteil des Anschubvorhabens sind intensive Forschungs- und Entwicklungsarbeiten in drei Kern-Technologiebereichen der Sprachtechnologie:

- Sprachbasierte Informationsextraktion und -fusion
- Dialogische Interaktion für Wissenszugang und -erwerb
- Sprachbasiertes Informationsmanagement und -retrieval

Diese F&E-Arbeiten wurden an der Universität der Saarlandes durchgeführt, um theoretische und praktische Ergebnisse aus den zahlreichen auf langfristige Grundlagenforschung angelegten Projekten der Arbeitsgruppen der Antragsteller, sowie anderer einschlägiger Forschungsvorhaben des SFB 378, des Graduiertenkollegs "Sprachtechnologie und Kognitive Systeme" und der Fachrichtungen Computerlinguistik und Informatik aufzunehmen und in Technologieverbesserungen umzusetzen.

I.2 Voraussetzungen der Vorhabensdurchführung

Die Sprachtechnologie wird heute unbestritten als eine Schlüsseltechnologie für die Zukunft der IT-Industrie, für den Ausbau der Informationsgesellschaft und für die technologische Infrastruktur der Wissensgesellschaft gesehen. Die Sprachtechnologie ist der Sammelbegriff für eine ganze Klasse von einzelnen Technologien, die durch algorithmisiertes Wissen über die Eigenschaften menschlicher Sprachen auf den Umgang mit Texten oder gesprochenen Äußerungen spezialisiert sind. Sie ermöglichen eine Fülle von Anwendungen, die von Diktiersoftware bis zu Dialogsystemen, von der Indizierung und Navigation auf dem WWW bis zur maschinellen Übersetzung reichen. Die wissenschaftlichen Grundlagen dieser Technologie kommen zu einem großen Teil aus der Computerlinguistik, aber zu geringeren Teilen auch direkt aus der Informatik, den Sprachwissenschaften und der Akustik.

Es gibt weltweit einige Standorte, an denen gezielt versucht wird, die notwendige Breite in der Sprachtechnologie herzustellen. Das ist aber erst ansatzweise erfolgreich. In Pittsburgh (USA) an der CMU sind zwar die Sprachtechnologien in bemerkenswerter Breite vertreten, es fehlen aber die Sprachwissenschaften und die computerlinguistische Basisforschung. In Stanford (USA) hingegen ist die Grundlagenforschung in Linguistik und Informatik hervorragend ausgebaut, es mangelt aber an Sprachtechnologien, die man höchstens in einigen Industriefirmen in der Umgebung findet. Nur an der Universität von Edinburgh (UK), wo sich in der computer-

linguistischen Grundlagenforschung der Bogen von den Kognitionswissenschaften bis hin zur theoretischen Informatik spannt, sind zumindest einige der wichtigsten Sprachtechnologien auch an der Universität vertreten. Deutschland kann zwar hervorragende Computerlinguisten und Sprachtechnologen aufweisen, nur sind diese über die Republik verstreut. Das bisher größte deutsche Forschungsvorhaben der Sprachtechnologie, der BMBF-Verbundprojekt Verbmobil, brauchte etwa zwanzig deutsche Partner, um alle relevanten Aspekte der Sprachverarbeitung gut abzudecken. Eine gewisse Breite findet sich lediglich in Saarbrücken, dem bedeutendsten deutschen Zentrum der Computerlinguistik und Sprachtechnologie, wo auch die Koordination des Verbmobil-Vorhabens angesiedelt war und derzeit auch die Projektleitung des MTI-Leitvorhabens SmartKom sitzt.

An den wenigen Standorten, die dem Idealbild des breit ausgestatteten Kompetenzzentrums nahe kommen, sind jeweils mehrere neue Sprachtechnologiefirmen entstanden, so z.B. in Pittsburgh, der Stanford-Umgebung, Edinburgh und Saarbrücken. Im Gegensatz zu den reiferen Ingenieurwissenschaften haben sich aber bisher in der Sprachtechnologie noch keine Zentren herausgebildet, die durch die umfassende Abdeckung der Teilgebiete und die enge Verbindung von wissenschaftlicher Grundlagenforschung, der Weiterentwicklung der Basistechnologien und Entwicklung realistischer Anwendungen zum Motor der internationalen technologischen Fortschritts und zum Werkzeug der Beschleunigung des Technologietransfers werden konnten.

Wie ständige Anfragen klar gezeigt haben, gibt es für ein solches Zentrum einen dringenden Bedarf. Durch das Bereitstellen von F&E-Kompetenz ist ein solches Zentrum ein wichtiger Mittler zwischen der Anwendergemeinschaft und den Technologieentwicklern.

I.3 Planung und Ablauf des Vorhabens

Das Vorhaben COLLATE wurde am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) und an der Universität des Saarlandes durchgeführt. Das Gesamtvorhaben wurde vom DFKI koordiniert. Am DFKI waren die Forschungsbereiche Sprachtechnologie (Prof. Uszkoreit) und Intelligente Benutzerschnittstellen (Prof. Wahlster) beteiligt. An der Universität des Saarlandes waren die Fachbereiche Computerlinguistik und Informatik beteiligt. Die Forschung wurde in drei Arbeitsgruppen durchgeführt, die von Prof. Pinkal (Computerlinguistik), Prof. Uszkoreit (Computerlinguistik) und Prof. Wahlster (Informatik) geleitet wurden. Das Projekt ist in vier Arbeitspakete gegliedert:

AP 1: Sprachbasierte Informationsextraktion und –fusion

Die Ergebnisse sind dieses Arbeitspakets sind in Abschnitt II.1.1 ausführlich dargestellt.

AP 2: Dialogische Interaktion für Wissenszugang und –erwerb

Die Ergebnisse dieses Arbeitspakets sind den Abschnitten II.1.2.3 (Dialogmodellierung für Informationszugriff) und II.1.5 (Multimodale Dialogarchitektur und Benutzergruppenerkennung) ausführlich dargestellt.

AP 3: Sprachbasiertes Informationsmanagement und –retrieval

Die Ergebnisse dieses Arbeitspakets sind den Abschnitten II.1.2 (NLP-Framework für Informationsmanagement), II.1.4 (Web-Korpora), und II.1.3 (Question Answering) ausführlich dargestellt.

AP4: Technologie-, Infrastruktur- und Projektkoordination

In diesem Arbeitspaket wurde die Zusammenarbeit mit dem DFKI und mit externen Partnern koordiniert, und der Austausch von Ressourcen und Technologien durchgeführt. Außerdem wurde die technische Infrastruktur (Linux-Cluster, Dialoglabor, Usability-Labor, Tonstudio, Gestenerkennungslabor) spezifiziert, beschafft, installiert und in Betrieb genommen.

Ein wesentliches F&E-Ergebnis des Projekts ist eine umfassende Informationsextraktions-Plattform mit einem Grammatikformalismus, Entwicklungswerkzeugen, Laufzeitsystem, Evaluationswerkzeugen, multilingualen Ressourcen, multilingualen Grammatiken, sowie Entwicklungs- und Testdaten. Forschungsergebnisse im Bereich von NLP-Methoden für Informationsmanagement sind ein Framework mit neuen und robusten Methoden für Chunking und topologisches Parsing, semantische Analyse und Annotation auf Basis von FrameNet und die praktische Anwendung eines Informationszustandsmodells (information state approach) für Dialogmanagement. Im Bereich von Question-Answering-Systemen für offene Domänen wurden erhebliche Verbesserungen in Bezug auf Geschwindigkeit, Robustheit und Skalierbarkeit eines existierenden Question-Answering-Systems erzielt. Die Wissensbasis des Systems wurde um acht Jahrgänge von CNN-Nachrichten erweitert. Wir haben Methoden und Werkzeuge entwickelt, um Web-Korpora mit umfangreichen Metadaten über Dokumente und Hyperlinks zu akquirieren und zu verwalten. Forschungsergebnisse im Bereich von mobilen, multimodalen und modularen Schnittstellen sind eine fortgeschrittene Systemarchitektur mit einer Medienfusionskomponente und effektive Algorithmen für die Erkennung von Benutzergruppen verschiedenen Alters und Geschlechts, anhand von akustischen Stimmmerkmalen.

Informationsextraktion

Die Anzahl der weltweit verfügbaren Informationen wächst ständig, so dass die Auswertung der Informationsquellen ohne leistungsfähige Werkzeuge nicht mehr machbar ist. Im Projekt COLLATE wurde ein System zur gezielten Extraktion von Informationen aus großen Textmengen entwickelt, das für neun europäische und asiatische Sprachen verfügbar ist. Mit dem System SProUT können gezielt Informationen über bestimmte Personen, Firmen oder Ereignisse aus Texten extrahiert und in eine Datenbank eingetragen werden.

SProUT wird eingesetzt zur automatischen Auswertung von Reisewarnungen für Krisengebiete, und für die Extraktion von Kundenmeinungen über Elektronik-Geräte und Kraftfahrzeuge aus Online-Diskussionsforen.

Question-Answering

Sucht ein Benutzer eine Antwort auf eine gezielte Frage, so muss er mit heutigen Suchmaschinen Stichwörter für eine Anfrage aussuchen, und dann mit großem Aufwand suchen, ob eines der gefundenen Dokumente die Antwort auf seine Frage enthält. In COLLATE wurde ein System weiterentwickelt und optimiert, das diese Schritte überflüssig macht, und direkt auf eine natürlichsprachliche Frage einen Satz mit der passenden Antwort liefert. Das System AnswerBus analysiert dazu die Frage des Benutzers, erstellt Stichwörter für eine Anfrage an eine Suchmaschine, und extrahiert Antworten aus den Ergebnisseiten. Auch das Abfragen von Nachrichtenquellen ist möglich. Das System wird täglich von einigen tausend Benutzern verwendet. Es wurde eine gesprochene Eingabe mit Spracherkennung realisiert, so dass demnächst Fragen auch telefonisch beantwortet werden können.

Mobile Assistenten für alle Benutzergruppen

Mit dem System M3I wurde eine Anwendung entwickelt, die Informations- und Navigationssysteme auf mobilen Geräten wie Handys oder PDAs verfügbar macht. Um eine komfortable Interaktion zu ermöglichen, wurden die Geräte mit Sprachein- und -ausgabe ausgestattet. Außerdem können sprachliche Eingaben mit Zeigegesten verbunden werden. Damit kann der Benutzer beispielsweise auf seinem Display auf einen Punkt auf dem Stadtplan zeigen und fragen „Wie komme ich dorthin?“ oder „Was ist das für ein Gebäude?“. Um die Geräte für möglichst viele Benutzergruppen verwendbar zu machen, optimiert sich die Spracherkennung automatisch für Geschlecht und Altersgruppe des Benutzers.

Internationaler wissenschaftlicher Beirat

Zur Sicherung der wissenschaftlich-technischen Qualität wurde ein wissenschaftlicher Beirat berufen, dem die folgenden international anerkannten Wissenschaftler angehören:

- Steven Bird, University of Melbourne und Linguistic Data Consortium
- Martin Kay, Stanford Univ. und XEROX PARC
- Sharon Oviatt, Oregon Graduate Institute of Science and Technology
- Donia Scott, University of Brighton
- Oliviero Stock, IRST Trento
- Hans Tillmann, BAS, Univ. München

Am 25. Oktober 2001 fand eine Sitzung des wissenschaftlichen Beirats statt. Dabei wurden die bisherigen Ergebnisse und der Arbeitsplan des Projekts vorgestellt und diskutiert. Die Empfehlungen des wissenschaftlichen Beirats wurden bei der weiteren Planung berücksichtigt. Ebenfalls am 25. Oktober 2001 fand die offizielle Eröffnung des Projekts statt, bei der hochrangige Vertreter aus Wissenschaft, Wirtschaft und Politik anwesend waren. Dabei wurden die Ziele und wissenschaftlichen Ansätze des Projekts präsentiert.

Am 7. Februar 2003 fand die Sitzung des wissenschaftlichen Beirats (Dr. Oliviero Stock, Prof. Dr. Hans Tillmann, Prof. Donia Scott, Prof. Dr. Martin Kay) statt. Dabei wurden Ergebnisse und Arbeitspläne vorgestellt. Von Seiten des wissenschaftlichen Beirats wurde die hohe Qualität der Projektarbeit und die gelungene Verbindung von innovativen Anwendungen und theoretischer Grundlagenarbeit gelobt. Das Projekt wurde als Modell für größere Aktivitäten, z.B. im europäischen Rahmen, bezeichnet.

I.4 Wissenschaftlicher und technischer Ausgangsstand

Das Ziel des Informationsmanagements besteht darin, das Informationsbedürfnis von Organisationen zu erfüllen. Die praktische Aufgabe ist daher die Verwaltung und Nutzbarmachung von sehr großen Informationsmengen. Diese soll dem einzigen Zweck dienen, die Information den Entscheidungsträgern genau dann zu liefern, wenn sie benötigt wird, und sie überdies so zu präzisieren, dass sie den Nutzer bei seiner Entscheidung effizient und effektiv unterstützt. Im folgenden wird der wissenschaftlich-technische Ausgangsstand in Einzelbereichen des natürlichsprachlichen Informationsmanagements dargestellt.

Informationsextraktion

Im Bereich der Informationsextraktion gab es vor allem Aktivitäten im englischsprachigen Bereich, wobei die technologische Entwicklung durch vergleichende Technologieevaluation im Rahmen der Message Understanding Conference (MUC) vorangetrieben wurde. Vergleichbare Evaluationen und Testdaten für die deutsche und andere europäische Sprachen waren nicht verfügbar.

In Saarbrücken gab es aus früheren Projekten eine Vielfalt von Forschungssystemen (SMES, MESON; SPPC), Ressourcen und Testdaten. Es fehlte jedoch eine benutzerfreundliche Entwicklungsplattform mit einem ausdrucksmächtigen Grammatikformalismus, Evaluationswerkzeug, sowie ein robustes, effizientes Laufzeitsystem.

Es gab nur Informationsextraktionssysteme für einzelne Sprachen, insbesondere für das Englische, jedoch noch keine Methodologie zur Entwicklung von parallelen Grammatiken für mehrere Sprachen, die eine systematische Wiederverwendung von Ressourcen (Ortsnamen, Personen- und Firmennamen) und Regelsystemen unterstützt, und damit die effiziente Entwicklung von Grammatiken für neue Sprachen ermöglicht.