Speech Understanding for Spoken Language Systems: Portability Across Domains and Languages

Zur Erlangung des akademischen Doktorgrades

Doktors der Ingenieurwissenschaften

von der Fakultät für Informatik der Universität Karlsruhe

(Technische Hochschule)

genehmigte

Dissertation

von

Wolfgang Minker

aus 89081 Ulm/Donau, Deutschland

Tag der der mündlichen Prüfung: 19. Dezember 1997

Hauptreferent: Prof. Dr. rer. nat. Alexander Waibel
Koreferenten: Dr. Lori Lamel
Prof. Dr. rer. nat. Alfred Schmitt

Abstract

This thesis investigates the problem of automatic natural language understanding for spoken language systems. The proposed parsing method is sufficiently general and flexible so as to be easily ported to different applications, domains and human languages.

Spoken language systems support unconstrained human-machine communication. They combine primary component technologies (such as speech recognition, natural language understanding and dialog processing) to understand the meaning of an input utterance. Natural language generation and/or speech synthesis are required to build end-to-end systems which accomplish some given task.

Today's state-of-the-art rule-based methods to natural language understanding provide good performance in limited applications for specific languages. However, the **manual** development of an understanding component using specific rules is costly as each application and language requires its own adaptation or, in the worst case, a completely new implementation. In order to address this cost issue, statistical modeling techniques are used in this work to replace the commonly-used hand-generated rules to convert the speech recognizer output into a semantic representation. The statistical models are derived from the **automatic** analyses of large corpora of utterances with their corresponding semantic representations. To port the semantic analyzer to different applications and languages it is thus sufficient to train the component on the application- and language-specific data sets as compared to translating and adapting the rule-based grammar by hand.

A stochastic method for natural language understanding was developed and applied to the following tasks and languages: the American ATIS (Air Travel Information Services), the French MASK (Multimodal-Multimedia Automated Service Kiosk) applications and the English Spontaneous Speech Task (ESST). The ATIS and MASK tasks deal with information retrieval for air and train travel, a domain of human-machine interaction. ESST deals with human-to-human interaction in which two people negotiate to schedule a meeting.

In ATIS, the corpora were semantically labeled by the rule-based component which was developed for the French language at the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (France). This same component was ported to English during the course of this thesis. For MASK, the semantic labels were obtained by integrating the stochastic component into the labeling process using bootstrapping and manual correction. For ESST, the model parameters were trained on a corpus of semantic tree-based represen-

tations which were produced by the natural language understanding component of JANUS, a spontaneous speech-to-speech translation system, in part developed at the University of Karlsruhe (Germany) and at Carnegie Mellon University (United States).

In direct comparison the stochastic data-driven parser is seen to outperform the rule-based method in terms of semantic accuracy and robustness. Furthermore, the semantic analyzer can be flexibly ported to new tasks, domains and languages. The strength of such a method is that the same software can be used regardless of the application and language. The stochastic models are trained on the specific data sets. The human effort in component development and porting is therefore limited to the task of data labeling, which is much simpler than the design, maintenance and extension of the grammar rules.

Zusammenfassung

Die vorliegende Doktorarbeit befaßt sich mit dem Problem des maschinellen Sprachverständnisses in einem Sprachverarbeitungssystem. Vorgestellt wird ein Verfahren zur semantischen Analyse, welches so allgemein und flexibel ausgelegt ist, daß es problemlos in neue Anwendungsbereiche und Sprachen übertragen werden kann.

Die praktische Sprachverarbeitungstechnik soll eine ungezwungene Verständigung zwischen Mensch und sprachgesteuerten Automaten ermöglichen. Derartige Systeme verwenden mehrere Grundsatzverfahren (das der Spracherkennung, des Sprachverständnisses und der Dialogbehandlung), um die Bedeutung eines gesprochenen Satzes automatisch zu erfassen. Zusätzlich werden Sprachgenerierung und/oder Sprachsynthese eingesetzt, um Gesamtsysteme zu entwickeln, die bestimmte Aufgaben erfüllen können.

Herkömmliche regelbasierende Verfahren zum maschinellen Sprachverständnis erzielen gute Resultate in begrenzten Anwendungsbereichen und Sprachen. Die **manuelle** Entwicklung einer solchen aus expliziten Regeln bestehenden Komponente ist jedoch kostspielig, denn jede Applikation und Sprache verlangt spezifische Anpassung oder schlimmstenfalls ein jeweils völlig neues System. Um diese Kostenfrage in der vorliegenden Arbeit zu erörtern, wird statistische Modellierung verwendet. Sie ersetzt die von Hand erstellten Regeln, welche die Ausgabe des Spracherkenners in eine semantische Darstellung umsetzen. Die Parameter der Wahrscheinlichkeitsmodelle leiten sich aus der **automatischen** Analyse großer Datenmengen von Benutzeranfragen und deren semantischen Darstellungen ab. Für den Einsatz einer semantischen Analysekomponente in unterschiedlichen Applikationen und Sprachen ist es daher ausreichend, diese Komponente mit anwendungs- und sprachenspezifischen Daten zu trainieren. Dies steht im Gegensatz zur manuellen Übersetzung und Anpassung einer regelbasierenden Grammatik.

In der vorliegenden Arbeit wurde ein stochastisches Verfahren zur semantischen Analyse entwickelt und auf die amerikanische ATIS (Air Travel Information Services), die französische MASK (Multimodal-Multimedia Automated Service Kiosk) sowie die amerikanische ESST (English Spontaneous Speech Task) Applikationen angewandt. ATIS und MASK liegen im Bereich der Mensch-Maschine Kommunikation und befassen sich mit Flugplan- und Zugfahrplanauskunft. ESST behandelt die Kommunikation von Mensch zu Mensch in Form von Terminabsprachen zwischen Verhandlungspartnern.

In ATIS wurden die semantischen Darstellungen zum Systemtraining aus den Benutzeranfragen von der am Laboratoire d'Informatique pour la Mécanique et les Sciences de

l'Ingénieur (Frankreich) für die französische Sprache entwickelten und im Rahmen der Arbeit ins Englische übertragenenen regelbasierenden Komponente automatisch erstellt. In MASK geschah dies unter Zuhilfenahme der stochastischen Komponente durch iteratives Etikettieren mit anschließend manueller Korrektur. Zur Berechnung der Modellparameter in ESST wurden Entscheidungsbäume von der semantischen Analysekomponente in JANUS, einem hauptsächlich an der Universität Karlsruhe (Deutschland) und an der Carnegie Mellon University (Vereinigte Staaten) entwickelten maschinellen Sprachübersetzer, erzeugt.

Eine direkte Vergleichsstudie zeigt, daß stochastische gegenüber konventionellen regelbasierenden Verfahren zum maschinellen Sprachverständnis bessere Ergebnisse erzielen, robuster sind und relativ schnell und unproblematisch in neue Anwendungsbereiche und Sprachen übertragen werden können. Die Stärken eines stochastischen Ansatzes liegen darin, daß durch die Wiederverwendung der Software der menschliche Arbeitsaufwand in Systementwicklung und Systemübertragung auf das Etikettieren von Daten begrenzt ist, mit denen die Analysekomponente dann automatisch die Modellparameter berechnet. Dies ist sehr viel einfacher als die Entwicklung, Aktualisierung und Erweiterung grammatikalischer Regeln.

Résumé

Cette thèse traite du problème de la compréhension automatique de la parole spontanée dans un système d'interaction vocale. Elle propose une méthode d'analyse sémantique suffisamment générale et flexible pour être facilement portée à différentes applications indépendamment de la langue.

Les technologies pour le traitement du langage parlé permettent une communication homme-machine flexible. Les systèmes, favorisant une interface naturelle, doivent combiner plusieurs traitements pour extraire le contenu sémantique d'un énoncé oral et accomplir une tâche définie : la reconnaissance de la parole, la compréhension du langage naturel, la gestion du dialogue ainsi que la synthèse vocale.

Dans une application limitée et dans une langue spécifique, les implémentations conventionnelles d'inférence de règles fournissent de bonnes performances. Néanmoins, le développement **manuel** d'un tel analyseur sémantique explicite est coûteux car chaque application et chaque langue nécessitent soit une adaptation, soit, dans le cas le plus extrême, une nouvelle implémentation. Afin de résoudre ce problème du coût, les techniques de modélisation stochastique générales et adaptables à d'autres applications et langues, peuvent se substituer aux méthodes par règles (symboliques) afin de produire une représentation sémantique à partir des phrases transcrites par le module de reconnaissance. Les modèles stochastiques résultent de l'analyse **automatique** d'un grand nombre de phrases provenant d'utilisateurs réels de l'application avec leur représentation sémantique correspondante. Ainsi, pour porter l'analyseur sémantique vers d'autres applications et langues, il suffit d'entraîner le module stochastique sur les données propres à l'application et à la langue, contrairement à la traduction et à l'adaptation manuelle dans le cas d'une grammaire par règles.

Dans le cadre de cette thèse une méthode stochastique pour la compréhension de la parole a été développée puis validée dans différentes applications et langues, dont ATIS (Air Travel Information Services) en anglais, MASK (Multimodal-Multimedia Automated Service Kiosk) en français et ESST (English Spontaneous Speech Task) en anglais. Les applications ATIS et MASK se situent dans le domaine de la communication homme-machine et concernent respectivement la demande d'informations sur les transports aériens et ferroviaires. ESST traite la communication homme-homme dans la situation d'une prise de rendez-vous entre deux interlocuteurs.

Pour ATIS, un système initial, en français, utilisant une méthode par règles a été réalisé

au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (France) et porté en langue anglaise, dans le cadre de cette thèse. Ce dernier système a permis d'étiqueter sémantiquement un corpus de phrases transcrites qui ensuite a été utilisé pour entraîner le module stochastique. Pour l'application MASK, le corpus sémantique a été établi à l'aide de l'analyseur stochastique en utilisant une méthode d'étiquetage itérative avec correction manuelle. Pour ESST, les paramètres du modèle ont été appris sur une représentation fondée sur des arbres, générés par l'analyseur sémantique de JANUS, système de traduction de parole spontanée développé principalement dans les Universités de Karlsruhe (Allemagne) et de Carnegie Mellon (États-Unis).

L'apport essentiel du travail présenté dans cette thèse est de montrer que pour des applications limitées, une méthode stochastique pour la compréhension de la parole spontanée est plus robuste. Lors d'une comparaison directe, cette méthode fournit de meilleurs résultats par rapport à un module conventionnel par règles. De plus, l'analyseur stochastique est réutilisable et peut être porté facilement vers d'autres applications, domaines et/ou langues. L'avantage réside dans le fait que l'effort humain se limite à l'étiquetage des données, qui sont ensuite utilisées lors de l'apprentissage des paramètres du modèle stochastique. Ceci est plus aisé que la conception, la maintenance et l'extension des règles de grammaire.

Acknowledgements

The work described in this thesis has been carried out at the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), France.

I would like to thank my advisors at LIMSI-CNRS, Jean-Luc Gauvain and Lori Lamel for suggesting the topic of this thesis, for introducing me into the problem of natural language understanding and for guiding me in my thesis work. I would also like to thank Alexander Waibel at University of Karlsruhe and Carnegie Mellon University for supporting my research so actively. Among other contributions, his suggestions for practical settings for the testing of my ideas were invaluable. I especially want to thank Alfred Schmitt for his interest in my work and his helpful suggestions.

I am particularly grateful to Françoise Néel, Samir Bennacef and Lin Chase (LIMSI-CNRS) who provided me with careful and well thought comments on my thesis draft. I would like to thank Samir Bennacef, Sophie Rosset, Christel Beaujard (LIMSI-CNRS) and Marsal Gavaldà (Carnegie Mellon University), who assisted me in a variety of experiments and evaluations reported in this dissertation. Thanks go also to Bernard Merienne and Michel Lastes (LIMSI-CNRS) who provided technical assistance with hardware-related problems.

Contents