In: KI – Künstliche Intelligenz, Heft 1, 2002

Disambiguierung durch Wissensfusion: Grundprinzipien der Sprachtechnologie

Wolfgang Wahlster
DFKI, wahlster@dfki.de

Der Beitrag ist eine Kurzfassung meines Ehrenvortrags anlässlich der Verleihung der Würde eines Ehrendoktors der Technischen Universität Darmstadt. Ausgehend von der Auflösung sprachlicher Mehrdeutigkeiten als ein Grundproblem der Sprachverarbeitung werden zunächst klassische Ansätze zur Disambiguierung skizziert. Danach werden aktuelle Arbeiten zur verzögerten Disambiguierung durch Unterspezifikation und zur wechselseitigen Disambiguierung multipler Eingabemodalitäten diskutiert.

1. Einleitung

Die Auflösung von Mehrdeutigkeiten sprachlicher Äußerungen (Fachausdruck: Disambiguierung) ist eines der Grundprobleme der maschinellen Sprachverarbeitung. Ein wesentliches Ergebnis der KI-Forschung auf dem Gebiet der Sprachtechnologie in den letzten 25 Jahren [4] ist, dass kein universeller Disambiguierungsalgorithmus existiert, sondern nur mithilfe der Kombination von Evidenzen aus verschiedenen Wissensquellen und durch die Fusion von Ergebnissen konkurrierender Verarbeitungspfade eine eindeutige Interpretation natürlichsprachlicher Dialogbeiträge erreicht werden kann. Mehrdeutigkeiten ergeben sich auf allen Stufen der Sprachverarbeitung, von der phonetischen, über die morphologische, lexikalische, syntaktische und semantische bis hin zur pragmatischen Ebene. Eines der Hauptprobleme der Disambiguierung in natürlichsprachlichen Dialogsysteme besteht darin, dass es rasch zu einer kombinatorischen Explosion der Lesarten kommt, wenn alle Alternativen einer Verarbeitungsebene zur nächsten Analysestufe weiterpropagiert werden. Bei der Verarbeitung gesprochener Sprache wird das Problem noch dadurch potenziert, dass durch die Spracherkennungsverfahren Wörtergitter mit alternativen Hypothesen für jedes gesprochene Wort entstehen, welche die Flut möglicher Lesarten noch weiter erhöhen. Eines der zentralen Probleme der Sprachtechnologie ist daher die Reduktion von Unsicherheit über die korrekte Interpretation einer sprachlichen Äußerung.

2. Selektionsrestriktionen

Eine der klassischen Methoden zur Disambiguierung ist die Verwendung von Selektionsrestriktionen und Weltwissen, wie folgendes einfache Beispiel veranschaulicht. Wenn ein Vater im künftigen Cyber-Restaurant dem Service-Roboter sagt: „Die Apfelschorle trinkt meine Tochter, die Weinschorle meine Frau“ dann führt die syntaktische Präferenz zunächst zu einer nicht sinnvollen Lesart wie *trinkt(Agens: Apfelschorle, Objekt: Tochter)* $\wedge$ *trinkt(Agens: Weinschorle, Objekt: Frau).* Wenn das System die Selektionsrestriktion *trinkt(Agens: Mensch, Objekt: Getränk*) auf das Weltwissen *Apfelschorle, Weinschorle* $\subset$ *Getränk* und *Tochter, Frau* $\subset$ *Mensch* anwendet, wird die richtige Lesart des Eingabesatzes als *trinkt(Agens: Tochter, Objekt: Apfelschorle)* $\wedge$ *trinkt(Agens: Frau, Objekt: Weinschorle)* selektiert. Selektionsrestriktionen haben wir bereits in dem ersten deutschsprachigen Dialogsystem HAM-RPM [1] verwendet. In VERBMOBIL [5] haben wir

eine verfeinerte Version dieser Grundidee auch für die Übersetzung eingesetzt. So wird das Wort *Termin* in drei verschiedenen Übersetzungen ins Englische übertragen abhängig von den Subsumptionsrelationen im Domänenmodell und von Selektionsrestriktionen. Dabei wird „Verschieben wir den *Termin*“ zu „Let's reschedule the *appointment*“ während „Schlagen Sie einen *Termin* vor“ zu „Suggest a *date*“ und „Da habe ich einen *Termin* frei“ zu „I have got a free *slot* there“ wird. VERBMOBIL war das erste Dialogübersetzungssystem, das neben Weltwissen auch den Kontext des vorausgehenden Satzes zur lexikalischen Disambiguierung und adäquaten Übersetzung nutzt. So wird nach „Nehmen wir dieses Hotel, ja?“ das Wort *Platz* in „Ich reserviere einen Platz“ mit *room* übersetzt, während nach „Gehen wir zum Abendessen“ das englische Wort *table* für *Platz* verwendet wird. Im Extremfall wird sogar die Systemuhr zur Äußerungszeit herangezogen, wenn vor 13 Uhr der Eingabesatz „Lassen Sie uns zusammen Essen gehen“ mit „Let's have lunch together“ und nach 17 Uhr als „Let's have dinner together“ übersetzt werden soll.

3. Unterspezifikation

Es ist nicht immer sinnvoll, eine eindeutige Bedeutungsrepräsentation sofort zu erzwingen, sondern oft ist es in Dialogsystemen ratsam, weitere Benutzereingaben abzuwarten, welche dann eine nachträgliche Disambiguierung einer zunächst noch mehrdeutigen Eingabe ermöglichen. Eine solche Wait-and-See-Strategie wird dann möglich, wenn zunächst unterspezifizierte Bedeutungsrepräsentationen erzeugt werden. Ein Eingabesatz mit klassischer Skopusambiguität wie „Einen Computer benutzen alle Informatikstudenten“ hat mindestens die zwei Lesarten (1) $\exists$ *x: computer (x)* $\forall$ *y: informatik-student benutzt (y,x)* oder (2) $\forall$ *y: informatik-student* $\exists$ *x: computer (x) benutzt (y,x).* Anstatt die Auswahl einer dieser Lesarten zu erzwingen, erzeugen moderne Dialogsysteme zunächst die unterspezifizierte Repräsentation {$\exists$ *x: computer (x),* $\forall$ *y: informatik-student}* *benutzt (y,x)*, welche ohne explizite Disjunktion die Skopusambiguität des Eingabesatzes reflektiert. Wenn dann der nächste Eingabesatz lautet „Das ist der Zentralrechner PDP-10“ (was vor 20 Jahren zu HAM-RPM-Zeiten noch der Fall war), dann wird nachträglich die Interpretation (1) gewählt. Wird der Dialog dagegen mit „Oft bringen sie ihr Notebook mit in die Vorlesung“ fortgesetzt, so ist die Lesart (2) zu wählen. Durch die kompakte Repräsentation mithilfe von Unterspezifikation wird eine kombinatorische Explosion der Lesarten vermieden und auf eine Auswertung aller Disjunktionen verzichtet. Es kann sogar eine nicht-monotone Diskurssemantik entstehen, wenn sich später im Dialog eine fehlerhafte Disambiguierung zeigt und für eine Reinterpretation zur unterspezifizierten Diskursrepräsentation zurückgekehrt werden muss. Die in den neunziger Jahren aufgekommene Verwendung unterspezifizierte Repräsentationen von Diskursbeiträgen behindert aber nicht die Inferenzfähigkeit der Systeme. So ist eine direkte Inferenz über der unterspezifizierten Formel {$\exists$ *x: computer (x),* $\forall$ *y: informatik-student}* *benutzt (y,x*) und $\forall$ *z: ki-student informatik-student(z)* möglich, die zu {$\exists$ *x: computer (x),* $\forall$ *y: ki-student}* *benutzt (y,x)* führt.

Bei der Dialogübersetzung in VERBMOBIL stellte sich heraus, dass oftmals ambiguitätserhaltende Übersetzungen sinnvoll sind. So bleibt die Ambiguität durch das PP-Attachment in dem Satz „Wir telephonierten mit Freunden aus Schweden“ (telephonieren wir aus Schweden oder wohnen die Freunde in Schweden?) in der englischen Übersetzung „We called friends from Sweden“ erhalten. Da der menschliche Dialogpartner die intendierte Lesart leichter findet, wäre in einem solchen Fall eine Disambiguierung der Eingabe durch VERBMOBIL überflüssig.

Durch Mechanismen zur Unterspezifikation bleibt die Mehrdeutigkeit beim semantischen Transfer in VERBMOBIL erhalten.

4. Multimodalität

Eine weitere wesentliche Neuerung in den letzten zehn Jahren ist die wechselseitige Disambiguierung multipler Eingabemodalitäten. Schon in unserem XTRA-Projekt hatte sich herausgestellt, dass eine kombinierte Sprach- und Gestikverarbeitung die Robustheit von Dialogsystemen erhöht [3]. Wenn man den akustischen Kanal mit einem Videokanal verbindet und die Analyseergebnisse einer Fusion unterwirft, wird die Disambiguierung oftmals erheblich vereinfacht. So erhöht sich die Robustheit von Spracherkennern bei gestörtem Sprachsignal und niedriger Worterkennungsrate durch gleichzeitiges Lippenlesen. Durch eine Kombination von Spracherkennung und Prosodieanalyse konnte in VERBMOBIL sowohl eine verbesserte lexikalische als auch syntaktische Disambiguierung von Dialogbeiträgen erreicht werden. In unserem Projekt SMARTKOM [6] werden die referenzsemantische Disambiguierung und die Aufmerksamkeitssteuerung durch die kombinierte Sprach- und Gestikerkennung wesentlich verbessert.

Es hat sich herausgestellt, dass eine Kombination von subsymbolischen und symbolischen Fusionsverfahren notwendig ist, um die Ergebnisse der Spracherkennung, der Prosodieerkennung, des Lippenlesens, der Gestikerkennung und der Mimikerkennung synergistisch zur Referenzauflösung und Disambiguierung einzusetzen. Das erstmals in XTRA [3] von uns entwickelte unifikations- und constraintbasierte Verfahren zur Kombination von Sprache und Gestik wurde von Johnston [2] weiter verfeinert. Inzwischen wurden auch Bayessche Netze und endliche Transduktoren zur symbolischen Wissensfusion eingesetzt, während auf der subsymbolische Ebene neuronale Netze und Hidden Markov Modelle dominieren (vgl. Abb. 1).

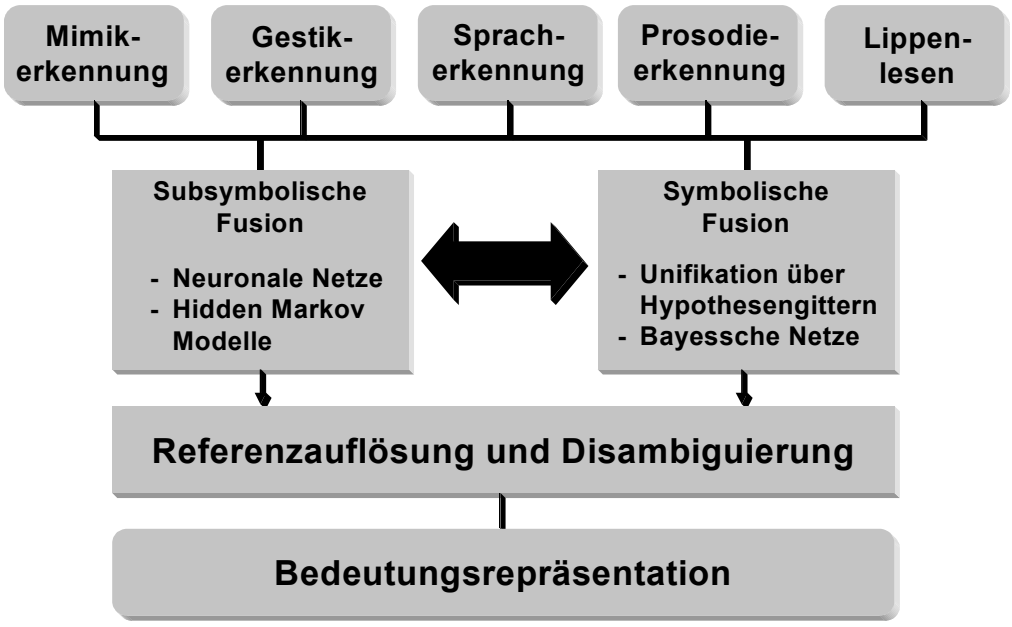


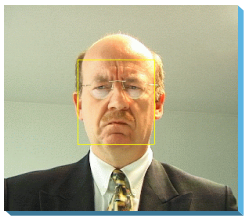
Abb. 1: Wechselseitige Disambiguierung multipler Eingabemodalitäten

Durch die Kombination von Spracherkennung und Mimikerkennung wird in SMARTKOM erstmals sogar die Erkennung von Ironie und Sarkasmus möglich. Abbildung 2 zeigt, dass die gleiche Spracheingabe des Benutzers „Echt toll“ je nach Gesichtsausdruck (ärgerlich versus neutral) zu einer ironischen oder nicht-ironischen Interpretation der Äußerung führt. Bei der Negation der Standardsemantik durch Erkennung von Ironie ergibt sich eine völlig andere Reaktion durch den Smartakus-Agenten.

Wichtig ist die Erkenntnis, dass nicht nur die Sprache durch die Gestik und Mimik disambiguiert werden kann, sondern auch umgekehrt eine mehrdeutige Geste oder ein unklarer Gesichtsausdruck durch die damit kombinierte sprachliche Äußerung auf eine eindeutige Bedeutungsrepräsentation abgebildet werden kann.

(1) Smartakus: Hier sehen Sie die Übersicht zum heutigen ZDF-Programm.

(2) Benutzer: Echt toll.



(3) Smartakus: Ich zeige Ihnen alternativ das Programm eines anderen Senders.

(2’) Benutzer: Echt toll.

(3’) Smartakus: Welche Sendungen wollen Sie aus dem ZDF-Programm sehen oder aufzeichnen?

Abb. 2: Fusion von Sprach- und Mimikerkennung

Während eines Dagstuhl-Seminars Ende Oktober 2001 zum Thema „Coordination and Fusion in Multimodal Interaction“ wurden von den international führenden Wissenschaftlern auf diesem aktuellen Gebiet für die nächsten acht Jahre Roadmaps für die Forschung entwickelt, die für den Leser im Netz unter der URL www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/ abrufbar sind.

Zusammenfassung

- Es gibt keinen universellen Disambiguierungsalgorithmus. Nur mithilfe der Kombination von Evidenzen aus verschiedenen Wissensquellen und durch die Fusion von Ergebnissen konkurrierender Verarbeitungspfade kann eine eindeutige Interpretation natürlichsprachlicher Eingaben gewonnen werden.
- Durch die wechselseitige Disambiguierung von Eingabemodalitäten (Sprache, Gestik, Mimik) sind multimodale Dialogsysteme erheblich robuster und effizienter als reine Sprachdialogsysteme.

- Die Unterspezifikation von Bedeutungsrepräsentationen erlaubt eine Verzögerung des Disambiguierungsprozesses im Dialog, bis ausreichende Information vorliegt, ohne notwendige Inferenzen zu blockieren.
- Bei der maschinellen Dialogübersetzung ist in vielen Fällen eine Disambiguierung des quellsprachlichen Ausdrucks gar nicht notwendig, wenn die Mehrdeutigkeit in der Zielsprache erhalten bleibt und ggf. vom Sprecher sogar intendiert war.

Literatur:

[1] v. Hahn, Walther, Hoeppner, Wolfgang, Jameson, Anthony, Wahlster, Wolfgang (1980): The Anatomy of the Natural Language Dialogue System HAM-RPM. In: Bolc, L. (ed.): Natural Language Based Computer Systems. München: Hanser/ Macmillan, p. 119-253.

[2] Johnston, Michael (1998) Unification-based Multimodal Parsing. In: Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics. (COLING-ACL 98), p. 624-630.

[3] Wahlster, Wolfgang (1991): User and Discourse Models for Multimodal Communication In: Sullivan, J. W. and Tyler, S. W. (eds.): Intelligent User Interfaces, New York : ACM Press, p. 45 - 67.

[4] Wahlster, Wolfgang (2000): Pervasive Speech and Language Technology. In: Wilhelm, R. (Ed.): Informatics - 10 Years Back, 10 Years Ahead. Berlin, Heidelberg, New York: Springer, Lecture Notes in Computer Science, Vol. 2000, p. 274 - 293.

[5] Wahlster, Wolfgang (2000) (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Berlin, Heidelberg, New York: Springer.

[6] Wahlster, Wolfgang., Reithinger, Norbert, Blocher, Anselm (2001): SmartKom: Towards Multimodal Dialogues with Anthropomorphic Interface Agents. In: Wolf, G., Klein, G. (eds.), Proceedings of International Status Conference "Human-Computer Interaction", Berlin: DLR, October 2001, p. 23 - 34.