4.6 Nicht-sprachliches Wissen

Kai-Uwe Carstensen

Nicht-sprachliches Wissen ist eine Untiefe im Meer computerlinguistischer Fragestellungen, die oft, insbesondere von formalen Semantikern, entweder weiträumig umschifft oder aber mit eher flachen Booten überquert wird. In diesem Abschnitt wird die Rolle nicht-sprachlichen Wissens als Ressource für die Computerlinguistik erläutert und es werden einige in diesem Bereich auftretende Probleme vorgestellt. Da auf elaborierte Darstellungen und platzgreifende Beispiele verzichtet werden muss, sollen vor allem die grundlegenden, für die Computerlinguistik relevanten Aspekte nicht-sprachlichen Wissens und seiner Repräsentation angesprochen werden.

4.6.1 Die Relevanz nicht-sprachlichen Wissens für die CL

Nicht-sprachliches Wissen ist aus folgenden Gründen essentiell für die Computerlinguistik:

- Sprachliche Zeichen weisen grundsätzlich sowohl eine Form- als auch eine –nicht-sprachliche– Inhaltsseite auf. Aus diesem Grund profitiert die Modellierung der einen Seite von den Erkenntnissen über die andere (dies gilt für beide Richtungen). Siehe hierzu als Beispiel *Lang et al.* 1991.

- Nicht-sprachliches Wissen ist zentraler Bestandteil intelligenter Systeme (daher: „wissensbasierte“ Systeme). Es wird zur Kategorisierung sensorischen Inputs, zur Problemlösung, zur Handlungsplanung und Kommunikation benötigt. *Konzeptuelle* Repräsentationen vermitteln zwischen Wahrnehmung, Handlung und Sprache.

- Das Fehlen von Wissen über die Welt führt(e) in den Anwendungssystemen der Künstlichen Intelligenz (KI), den Expertensystemen, zur so genannten „Zerbrechlichkeit“ (*brittleness*): Schon die geringsten Abweichungen von den vorgegebenen Eingabemustern (insbesondere auch von Menschen als „einfach“ eingestufte Fragen) führ(t)en zu Fehlern und Systemabstürzen. Hieraus enstand das Desiderat allgemein und wieder- verwendbarer Wissensressourcen (sog. *(common sense) knowledge sharing and reuse*).

- Komplexe sprachverarbeitende Systeme sind ohne nicht-sprachliches Wissen undenkbar: im Verlauf des *Textverstehens* müssen Textrepräsentationen mit Hintergrundwissen verrechnet werden (z.B. für die Auflösung von Ambiguitäten und bei der Präsuppositionsrechtfertigung); nichtsprachliche Wissensstrukturen sind Grundlage und Ausgangspunkt für die *Sprachgenerierung* (s. Unterkapitel 5.12); in der *maschinellen Übersetzung* (s. Unterkapitel 5.13) werden konzeptuelle Repräsentationen als Interlingua verwendet.

4.6.2 Wissen und Wissensrepräsentation

Kerngebiet der Beschäftigung mit nicht-sprachlichem Wissen ist der Bereich der **Wissensrepräsentation** innerhalb der KI bzw. der Kognitionswissenschaft. Interessanterweise waren es vor allem sprachlich orientierte Ansätze, die die Notwendigkeit und das Potential der Repräsentation nicht-sprachlichen Wissens aufgezeigt haben (z.B. *Quillian* 1968, *Schank* 1975), sowie das bekannt-berüchtigte ELIZA-Programm Joseph Weizenbaums als Karikatur eines Systems, das gerade nicht über Wissensrepräsentationen verfügt.

„Wissensrepräsentation" bezeichnet einerseits die Realisierung abstrakten **Wissens** in einem konkreten (physikalischen) System (also Mensch oder Maschine) und andererseits die Strukturen, die sich aus der Interaktion eines vorstrukturierten informationsverarbeitenden Systems mit seiner Umwelt (→ Lernen) ergeben. Das Verhältnis von Wissen zu dessen Repräsentation ist am markantesten in der sogenannten *Knowledge Representation Hypothesis* ausgedrückt:

> Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but causal and essential role in engendering the behavior that manifests that knowledge. (*Smith* 1982, S. 33).

Den „structural ingredients" entsprechen in *symbolischen Ansätzen* zur Wissensrepräsentation Systeme von Symbolen mit einem jeweils spezifischen Bedeutungsgehalt, wobei die Symbole als in irgendeiner Weise physikalisch realisiert aufgefasst werden (sog. *Physical symbol system hypothesis*, s. *Newell und Simon* 1976). Ein Beispiel hierfür sind Symbole für Konzepte, denen in der realen Welt Einzeldinge oder Mengen solcher Dinge entsprechen. Die Frage jedoch, wie diese Symbolsysteme genau in den Erfahrungen bzgl. der Umwelt verankert sind, ist als das „symbol grounding problem" (*Harnad* 1990) bekannt geworden. Dem gegenüber stellen *subsymbolische* bzw. *konnektionistische Ansätze* das Verkörpertsein (embodiment) und die Situiertheit neuronaler Netze, die von vornherein auf einer Sensor/Input-Aktor/Output-Korrelation beruhen und in denen sich ein von außen zugeschriebener Bedeutungsgehalt (z.B. „Dies ist das Konzept für X") aus den Aktivationsmustern einer Vielzahl von Neuronen/Einheiten ergibt.

Beide Ansätze haben ihre Stärken und Schwächen und sind gegenwärtig daher als komplementär zueinander anzusehen. Beispielsweise erfassen subsymbolische Ansätze sehr viel besser die graduellen Unterschiede, kontextuellen Abhängigkeiten und impliziten Zusammenhänge in eng begrenzten Anwendungsbereichen. Symbolische Ansätze bleiben jedoch vorerst für die Entwicklung komplexer natürlich-sprachlicher Systeme besser geeignet. Dies gilt insbesondere für die Erstellung umfangreicher Ressourcen nicht-sprachlichen Wissens.

4.6.3 Aspekte derWissensrepräsentation

Allgemeine Aspekte

Kern der Wissensrepräsentation ist die Darstellung der im Folgenden aufgeführten generischen Wissensrepräsentationskonstrukte: **Konzepte** (dt. Begriffe) als Repräsentanten von Entitäten der Welt (zu unterscheiden sind hier Klassenkonzepte (→ generisches Wissen über Dinge) und Individuenkonzepte (→ Wissen über Einzeldinge); **Attribute** als Repräsentanten der Eigenschaften solcher Entitäten; **Relationen** als Repräsentanten von Beziehungen zwischen Dingen; **Regeln** als Repräsentanten der Beziehungen zwischen Sachverhalten.Die Aufgabe der Wissensrepräsentation ist die formale Explikation dieser Aspekte, so dass alles relevante Wissen (je nach Anspruch eingeschränkt auf bestimmte Bereiche (*Domänen*) oder Verwendungszwecke) entweder direkt repräsentiert ist oder anhand von Schlussfolgerungen (*Inferenzen*) systematisch erschlossen werden kann. Hierbei stellen sich viele Detailfragen (z.B.: Welche Repräsentationskonstrukte entsprechen dem Ausdruck „ist ein“?), deren Beantwortung die Kenntnisse der kognitionswissenschaftlichen Disziplinen (u.a. Informatik, Linguistik, Psychologie, Philosophie) erforderlich macht.

Paradigmen

Verschiedene Sichtweisen darauf, wie sich aus solchen allgemeinen Wissensrepräsentationskonstrukten konkrete Wissensrepräsentationsstrukturen konstruieren lassen (und welche Prozesse darüber ablaufen sollen) haben zu unterschiedlichen Paradigmen der Wissensrepräsentation geführt. Den **semantischen Netzwerken** liegt die Idee der Vernetztheit konzeptuellen Wissens zugrunde. Ihren Ursprung hat diese Auffassung in den Arbeiten Quillians, der entsprechende Repräsentationen zur Berechnung der inhaltlichen Beziehung sprachlicher Elemente (daher: „semantische“ Netzwerke) verwendete. Den „Knoten“ („nodes“) des semantischen Netzwerks entsprechen die Konzepte, den „Kanten“ („links“) die vielfältigen Beziehungen zwischen ihnen. Entsprechend lassen sich in einem solchen Netzwerk „Nähe“ bzw. „Ferne“ von Konzepten über die Länge der Pfade verbindender Kanten verstehen und auch psychologisch relevante Prozesse wie den der Aktivationsausbreitung definieren. Das **Frame-Paradigma** (*Minsky* 1975) betont den objekt-orientierten und schematischen Aspekt der Wissensrepräsentation, wonach das relevante Wissen über eine Entität direkt an ihrem Stellvertreter verfügbar ist. Zentrales Konstrukt dieses Paradigmas ist das des Frames („Rahmen“) als Repräsentation schematischen Wissens über Entitäten der Welt (Stühle, Kindergeburtstage etc.). Frames sind im Wesentlichen Attribut-Wert-Paare, wobei die Werte („Filler“) der Attribute („Slots“) bei Fehlen genauerer Information durch *typische* Information („per Default“) instanziiert werden können, die ggf. überschrieben wird und die insgesamt den Prototyp eines Frames charakterisiert. Frames dienen einerseits der Klassifikation vorliegender Information und andererseits dem Inferieren weiterer Information. Sie können in hierarchischer Beziehung zueinander stehen (und somit *Taxonomien* darstellen), so dass Information an untergeordneten Frames „vererbt“ werden

kann. Das **Logik-Paradigma** betont die Uniformität der Darstellung von Wissen (in erster Linie mit Hilfe der Prädikatenlogik erster Stufe oder Varianten davon), insbesondere auch für die Anwendbarkeit allgemeiner Inferenzmechanismen (Theorembeweiser). Sein Nachteil besteht in der Strukturarmut logischer Repräsentationen. Das **(Produktions-)Regel-Paradigma** fokussiert auf den Aspekt der Steuerung des Verhaltens eines Systems durch die Anwendung von (Wenn-Dann-) Regeln auf jeweils aktuelle Daten.

Moderne Wissensrepräsentationssysteme wie z.B. LOOM (`http://www.isi.edu/isd/LOOM/LOOM-HOME.html`) stellen oft eine Mischung aus diesen Paradigmen dar.

Unterscheidungen

Während der *Inhalt* nicht-sprachlichen Wissens auf der *konzeptuellen* Ebene eines Wissensrepräsentationssystems spezifiziert wird, wird dessen *Form* bzw. *Struktur* nach *Brachman* 1979 auf der *epistemologischen* Ebene determiniert. Hierzu gehören Kanten, die die hierarchischen Beziehungen zwischen Konzepten (Subkonzept-Superkonzept, Individuenkonzept-Klassenkonzept), oder solche, die die „Rollen“ von Konzepten (Relationen, Attribute), darstellen. Diese beiden Ebenen der Repräsentation wurden in frühen Systemen oft mit drei weiteren vermischt: der *implementatorischen* (Kanten als Pointer), *logischen* (Kanten als Junktoren) und *linguistischen* Ebene (Kanten als sprachliche Relatoren).

Insbesondere den frühen semantischen Netzwerken mangelte es generell an formaler Klarheit ihrer Repräsentationskonstrukte, vor allem der Kanten (vgl. den Titel von *Woods* 1975, „What's in a link“). Auch wenn der Einsatz der Netzwerke teilweise zu beeindruckenden Ergebnissen führte (wie die Verwendung der *konzeptuellen Dependenzstrukturen* Roger Schanks für das Textverstehen), blieb ein wesentlicher Teil ihrer Bedeutung oft in ihrem Verarbeitungsmechanismus versteckt. Seither gilt es als Maxime, Sprachen zur Beschreibung von Repräsentationskonstrukten mit einer entsprechend expliziten Semantik zu versehen (daher auch der Terminus **Beschreibungslogik** (*description logic*) für die Spezifizierung von (bestimmten) Wissensrepräsentationsformalismen), die einen Bezug der Konstrukte zu ihrer Interpretation in der Welt herstellt.

Die Interpretationsaufgabe setzt wiederum Kenntnis über die Welt an sich voraus, wie sie seit langem in der philosophischen Disziplin der *Ontologie* (d.h. „Lehre vom Seienden“) gesammelt wird. In der Wissensrepräsentation wird unter einer **Ontologie** im engeren Sinne nach *Gruber* 1995 eine *explizite Spezifikation einer gemeinsamen Konzeptualisierung* verstanden, mit der nichts Anderes als eine formale intersubjektive sprachunabhängige Repräsentation der Welt gemeint ist, die von den Aufgaben, Zielen, Handlungen, Einstellungen etc. eines Agenten sowie von den Spezifika einzelner Situationen abstrahiert.

Eine Ontologie beschränkt sowohl die Primitive der konzeptuellen Ebene bzgl. ihrer möglichen Interpretationen, als auch die möglichen Optionen auf der epistemologischen Ebene. Zwischen diesen beiden Ebenen ist deshalb nach *Guarino* 1995 eine eigenständige *ontologische* Ebene anzunehmen. Innerhalb des ontologischen Wissens werden bereichsspezifische (*domain ontologies*, *lower-level on-*

tologies) von allgemeinen Aspekten (*top ontology*, *upper-level ontology*) unterschieden, wobei insbesondere der upper-level ontology als wiederverwendbarer Ressource eine besondere Bedeutung zukommt.

In einem weiter gefassten Sinn hingegen versteht man unter Ontologien allgemein die Wissensressourcen eines wissensbasierten Systems, weshalb beispielsweise auch lexikalisch-semantische Ressourcen wie WordNet (s. Unterkapitel 4.3) zum Teil als Ontologien bezeichnet werden.

In Bezug auf das konzeptuelle Wissen muss zwischen dem, was es gibt (*immanentes Wissen*), und dem, was (in einem bestimmten Kontext) der Fall ist (*situatives Wissen*), unterschieden werden. Das immanente Wissen umfasst sowohl die notwendigen Merkmale eines Konzepts (z.B. dass ein Elefant ein Tier ist) wie auch die das *Defaultwissen* ausmachenden (z.B. dass Elefanten vier Beine haben). Gleichzeitig muss es möglich sein, Konzepte zu bilden, die den Defaultannahmen widersprechen (z.B. DREIBEINIGER-ELEFANT), u. a. um spezifische Aussagen/Assertionen zu ermöglichen („Clyde ist ein dreibeiniger Elefant“).

Hieraus resultiert die Unterscheidung einer terminologischen Komponente („T-Box“) und einer assertionalen, Faktenwissen repräsentierenden Komponente („A-Box“) in Frame-basierten Wissensrepräsentationssystemen wie KL-ONE (*Brachman und Schmolze* 1985). Üblicherweise werden T-Box und A-Box, ggf. zusammen mit einer Regelkomponente, als die **Wissensbasis** eines entsprechenden Systems bezeichnet. **Wissensbasierte Systeme** verfügen neben der Wissensbasis über einen **Inferenzmechanismus**, der die Wissensbasis manipulieren kann (zum Aufbau einer Wissensbasis s. *Brachman et al.* 1990).

4.6.4 Wissensrepräsentation für die CL

Probleme

Die Verwendung nicht-sprachlichen Wissens in natürlich-sprachlichen Systemen ist grundsätzlich nicht-trivial, da die existierenden Ressourcen weder ausgereift noch generell mit sprachlichen Komponenten kompatibel (d.h. für natürlich-sprachliche Zwecke geeignet) sind. Im Einzelnen lassen sich die folgenden Probleme nennen:

Die notwendige **Trennung nicht-sprachlicher und sprachlicher Konzepte** wird nicht immer strikt eingehalten. Zum Einen werden in der Wissensmodellierung sprechende Bezeichner (d.h. sprachliche Symbole) verwendet, was zumindest potentiell (wohl aber sogar faktisch) zu einem sprachlichen und auch kulturspezifischen „Bias“ führt (wodurch insbesondere bei Interlingua-basierter Übersetzung Probleme auftreten können). Zum Anderen werden sprachliche Ressourcen wie WordNet nicht selten als Ontologien im engeren Sinne verwendet. Durch Anwendung ihrer OntoClean-Methodologie zeigen *Gangemi et al.* 2002, dass WordNet für diesen Zweck nicht geeignet ist, da dessen Konzeptstruktur einigen grundlegenden ontologischen Wohlgeformtheitsbedingungen nicht genügt.

Die **Nicht-Berücksichtigung sprachlich relevanter Differenzierungen** in Ontologien bedeutet deren Unbrauchbarkeit für computerlinguistische Zwecke. Entsprechend müssen die ausschließlich von spezifischem Domänenwissen abstra-

hierenden Top-Ontologien um sprachlich relevante Konzepte erweitert werden. Ein Vorschlag für eine entsprechende upper-level Ontologie ist das *Generalized Upper Model* von *Bateman et al.* 1994.

Grundlegende Probleme der Wissensmodellierung, wie z.B. die **Bestimmung eines einheitlichen Inventars an Repräsentationsprimitiven**, sind immer noch nicht gelöst. So zeigt ein Vergleich von Wissensrepräsentationssystemen eine erschreckende Uneinheitlichkeit schon bei grundlegenden Repräsentationskonstrukten.

Ein besonderes Problem ergibt sich aus der weitgehenden **Arbitrarität konzeptueller Slots** im Frame-Paradigma. Zum Beispiel ist ‚`bodyPartsRequiredOfPerformer`' im System **Cyc** nach *Lenat und Guha* 1989, S. 192 ein Slot eines ‚`EatingPopcorn`'-Frames (mit den Werten ‚`Teeth, Mouth, Throat, Stomach, Brain`'[!]). Slots wie diese verstecken komplexe Strukturzusammenhänge hinter sprachlich eingängigen Bezeichnungen. Deren Beziehung zum restlichen Wissen (bzw. deren Semantik) ist entweder nicht vorhanden oder nur durch ein komplexes „Slot-Bookkeeping" herzustellen. Probleme wie diese haben übrigens zur Aufgabe der Frame-Basiertheit in Cyc zugunsten einer logisch verteilten Repräsentation (sog. „knowledge soup") geführt. Wie *Mahesh et al.* 1996 in ihrer Analyse der NLP-Tauglichkeit von Cyc zeigen, hat jedoch die daraus resultierende Unstrukturiertheit der Repräsentation ebenfalls negative Konsequenzen für deren Verwendbarkeit in natürlich-sprachlichen Systemen.

Ein immer noch bestehendes praktisches Problem ist die **mangelhafte Verfügbarkeit** (vor allem kommerziell entwickelter) vorhandener Ressourcen.

Geschichte

Nach dem Aufkommen der semantischen Netzwerke (Ende der 60er Jahre) und des Frame-Paradigmas (Anfang der 70er Jahre) setzte eine „Logifizierung" der Wissenrepräsentationsformalismen ein, aus der verschiedene (Default-)Logiken sowie die Beschreibungslogiken (auch: *terminologische Logiken*) resultieren. Die deutlich werdende „Zerbrechlichkeit" der wissensbasierten Systeme (vor allem der Expertensysteme) führte 1984 zu dem Projekt Cyc (von „Encyclopedia") (*Lenat und Guha* 1989), in dem, ausgehend von der Entwicklung einer top ontology, eine umfassende, „common sense knowledge" repräsentierende Wissensbasis entwickelt werden sollte (und immer noch wird). Die 90er Jahre sahen ein rasant ansteigendes Interesse an Ontologien, insbesondere auch unter dem Aspekt der Wiederverwendbarkeit von Wissen im oder für das WWW. In diesem Rahmen lässt sich auch eine Renaissance semantischer Netzwerke für die Klassifikation und das selektive Retrieval von Web-Dokumenten unter dem Stichwort **Topic maps** beobachten, mit all ihren assoziierten Problemen (derer sich ihre Erfinder nicht unbedingt bewusst sind).

Perspektiven

Die Nutzung nicht-sprachlichen Wissens als computerlinguistische Resssource wird, nicht zuletzt durch den bislang ausgebliebenen Erfolg von Cyc, in der näch-

sten Zeit voraussichtlich vor allem pragmatisch ausgerichtet sein. Dies bedeutet in erster Linie die Aufgabe der Erwartung einer *perfekten* Gesamt-Ontologie, einer *idealen* Interlingua, einer *alles abdeckenden* Top-level Ontologie oder einer *allumfassenden* Wissensbasis. Stattdessen werden existierende Ressourcen trotz partieller Inkompatibilitäten zusammengefügt („ontology merging") wie z.B. bei der Entwicklung der *Sensus*-Ontologie (*Knight und Luk* 1994).

Aufgrund der festgestellten Lücken und Qualitätsunterschiede in der Wissensbasis von Cyc schlägt Mahesh (*Mahesh und Nirenburg* 1995,*Mahesh* 1996) im Rahmen der auf maschinelle Übersetzung ausgerichteten Entwicklung der *Mikrokosmos*-Ontologie einen *situierten* Ansatz der Ontologie-Konstruktion vor, bei dem die Wissensmodellierung insbesondere von der Aufgabe und dem frühen praktischen Einsatz der Wissensbasis geleitet wird. Dabei werden sprachlich relevante Unterscheidungen einerseits von vornherein und andererseits so weit wie nötig bei der Ontologie-Konstruktion berücksichtigt, so dass die Erstellung einer ontologischen Interlingua einen *approximativen* Charakter erhält.

Komplementär dazu verfolgt Guarino eine formal-ontologisch gesteuerte Wissensmodellierung (*Guarino* 1995), indem er ontologische Meta-Eigenschaften als Constraints für wohlgeformte Ontologien verwendet. Zusammen mit den intensivierten informatischen Bestrebungen auf den Gebieten Standardisierung von Wissensrepräsentationsformalismen und Tool-Development (Ontologie-Browser etc.) lassen diese Entwicklungen erhebliche Qualitätsverbesserungen erwarten. Angesichts der Tatsache, dass auch der Ansatz Guarinos stark von linguistischen Überlegungen geprägt ist, sind es paradoxerweise die (Computer-)Linguisten, die Wesentliches im Bereich nicht-sprachlichen Wissens beitragen.

4.6.5 Literaturhinweise

Standardwerke der Einführung in die Wissensrepräsentation im Hinblick auf die Verarbeitung natürlicher Sprache sind *Sowa* 1984 und *Sowa* 2000. Eine lesbare Einführung mit vielen Beispielen ist *Reimer* 1991. Empfehlenswert im Hinblick auf Aktualität, Umfang und Detailreichtum sprachlich orientierter Wissensrepräsentation ist *Helbig* 2001. Interessierte finden die klassischen Papiere zur Wissensrepräsentation in *Brachman und Levesque* 1985. Ein Einstieg in das Thema description logics ist `http://dl.kr.org`.

John Bateman (`http://www-user.uni-bremen.de/~bateman/`) bietet eine Sammlung relevanter Information zu Ontologien und weiteren Aspekten der Wissensrepräsentation an, die insbesondere im Hinblick auf die Entwicklung von Grundlagen für das Semantic Web (s. Unterkapitel 4.7) relevant ist.