

Universität Osnabrück
Sprach- und Literaturwissenschaften
CL / KI
SS 2000
Kollokationen
Leitung: P. Ludewig und H. Gust

Lexikographie aus der Sicht von Computerlinguistik und Künstlicher Intelligenz

Britta Koch
Am Salzmarkt 3, WG 11
49074 Osnabrück
M.A. Computerlinguistik und Künstliche Intelligenz
CL/KI(6) / Philosophie(6) / Informatik(6)

Inhaltsverzeichnis

1	Einleitung	2
2	Das Lexikon und die KI	3
2.1	Wissensklassifikation	3
2.2	Wissensrepräsentation	4
3	Computerlinguistik und das Lexikon	6
4	Vergleich der beiden Texte	8
5	Fazit	10
A	Literaturverzeichnis	11

1 Einleitung

Die Lexikographie ist eine eigenständige Wissenschaft, aber auch Computerlinguistik und KI-Forschung bemühen sich um sie. Die Computerlinguistik, weil beim Verarbeiten von Sprache ein gutes Lexikon sehr wichtig ist, und die KI, weil sie bei der Erstellung von Ontologien und Wissensbasen für die Lexikographie wichtige Erfahrungen gemacht hat. Im Folgenden werde ich mich mit 2 solchen Fällen auseinandersetzen. Christoph Habel hat in seinem Text „Das Lexikon in der Forschung der Künstlichen Intelligenz“ die Sicht der KI beschrieben, wobei Bran Boguraev und Ted Briscoe in ihrer Einführung zum Buch „Computational Lexicography for Natural Language Processing“ sich mehr dem computerlinguistischen Standpunkt gewidmet haben.

Zunächst werde ich beide Texte kurz zusammenfassen und sie dann miteinander vergleichen. Am Schluß stelle ich dann beides in den größeren Zusammenhang des Seminars „Kollokationen“.

2 Das Lexikon und die KI

Habel stellt schon am Anfang klar, daß er mit dem Begriff „Lexikon“ nicht nur ein Wortformenverzeichnis, sondern eine Repräsentation der in der natürlichen Sprache verwendeten Begriffe meint. Für gute Lexika gibt es viele Anwendungsgebiete, aber oft beeinflußt auch das Anwendungsgebiet die Form und den Umfang der Lexika. Zwei der Informatik entnommene Annahmen sind dem Autor wichtig: daß 1. sprachliche Prozesse informationsverarbeitend und 2. sprachliche Phänomene mithilfe sprachlicher Prozeduren zu erklären sind. Um aber das Ganze auf einem Computer laufen zu lassen, muß man es formalisieren, was eine Beschränkung sein kann. Weil kognitive Prozesse wissens- bzw. informationsbasiert sind, also durch Wissen gesteuert werden, müssen u. U. auch Weltwissen oder Erinnerungen in die Sprachverarbeitung einfließen.

2.1 Wissensklassifikation

In der Sprachverarbeitung muß es eine wohldefinierte Beziehung zwischen den Ausdrücken und Bedeutungsrepräsentationen geben. Die Produktion sowie das Verstehen natürlicher Sprache wird durch Vorwissen wie Erfahrungen, Erinnerungen etc. gesteuert - bei unterschiedlichem Vorwissen können unterschiedliche Bedeutungsrepräsentationen entstehen. Vor allem Tilgungsvorgänge und umgekehrt das Auffüllen von Lücken wird durch das Vorwissen des Systems gesteuert.

Habel teilt nun Wissen grob in drei Klassen ein: Er stellt semantisches dem episodischen Wissen gegenüber, und teilt das semantische in lexikalisches und enzyklopädisches Wissen auf. Episodisches Wissen bezieht sich auf einzelne Situationen oder autobiographische Erfahrungen, wohingegen lexikalisches Wissen das Wissen von Wörtern und Bezeichnungen und enzyklopädisches Wissen das Wissen um Kategorien und Einordnungsmöglichkeiten beinhaltet. Weil aber diese drei Arten des Wissens oft zusammenspielen, kann man sie nicht gut voneinander getrennt behandeln. Vor allem die Auflösung von Mehrdeutigkeiten beruht meist auf mehr als einer Wissensklasse. Danach zeigt Habel dieses Dilemma anhand der Sätze „Flying planes can be dangerous“ versus „Flying bees can be dangerous“ und „Müllers sahen die Alpen, als sie nach Süden flogen“ versus „Müllers sahen die Kraniche, als sie nach Süden flogen“, und macht einige Lösungsvorschläge, z. B. bei Verben die Agenten anzugeben, oder die Verwendung von Features zur Kennzeichnung von möglichen Objekten. Auch das Erkennen von temporalen und kausalen Bezügen ist abhängig vom Weltwissen - durch das Wort „weil“ erkennt der Leser eine Kausalbeziehung und formuliert eine Kette von Bezügen, die je nach Wissen unterschiedlich lang ist.

Schließlich erwähnt der Autor noch die Unterscheidung von prozeduralem und dekla-

rativem Wissen, also Wissen, wie und Wissen, daß. Diese Unterscheidung kann man auch bei der Programmierung von Systemen treffen - man kann entweder alle Daten in Listen verpacken, oder mit Prozeduren ausstatten. Zur Auswahl des Wissens im nächsten Schritt wird das Kontrollwissen verwendet. Es steuert außerdem die Suche und Inferenz des nötigen Wissens. Die beiden Vorgehensweisen prozedurale vs. deklarative Repräsentation von Wissen haben unterschiedliche Vor- und Nachteile. Bei deklarativen Wissenssystemen kann man flexibel und ökonomisch arbeiten, weil der Prozeßablauf nicht im Voraus geplant werden muß und das Kontrollwissen ist nur implizit vorhanden. Aber wenn man es übertreibt und alles als Daten betrachtet, wird diese Vorgehensweise ineffizient, weil man immer neu feststellen muß, welches Wissen relevant ist. Bei prozeduralem Wissen gibt man das Kontrollwissen explizit an und kann bei jeder Wissensseinheit einen Verweis auf relevantes Wissen geben. Idealerweise verbindet man die beiden Vorgehensweisen jedoch.

2.2 Wissensrepräsentation

In diesem Abschnitt stellt Habel einige Ansätze aus der Wissensrepräsentation vor. Zunächst erklärt er semantische Netze, netzartige Ontologien, die mit *is-a*- und *has-prop*-Kanten arbeiten. Bei diesen ist die Organisation wichtig. Innerhalb des Netzes gibt es eine Klassenhierarchie und Vererbungsmöglichkeiten, wobei Vererbung auch blockiert werden kann. Er erläutert das Konzept des „semantischen Abstands“ - eine Eigenschaft wird über mehrere Kanten vererbt, und ist deshalb nicht so präsent wie unmittelbar vererbte Eigenschaften. Um Redundanz zu vermeiden, sollte man Konzepte am höchstmöglichen Knoten zuordnen, aber Explizitheit, Relevanz oder Häufigkeit einer Eigenschaft führen manchmal zur redundanten Zuordnung - hier wird die Debatte zwischen Speicherungsökonomie und Verarbeitung- bzw. Inferenzökonomie erwähnt. Außerdem gibt es Konzepte, die Individuen entsprechen, sowie generische Konzepte, die durch gewisse charakteristische Eigenschaften bestimmt werden.

In diesem Zusammenhang erwähnt Habel auch die Conceptual-Dependancy-Theorie von Schank, und gibt seine Beispiele in dieser Notation wieder. Mithilfe dieser Skizzen findet er vier Bedeutungen des Wortes „fliegen“, so daß er die oben erwähnten Beispielsätze unterscheiden kann. Allerdings erwähnt er auch das Problem, daß man sich bei semantischen Netzen in Details verstricken kann, oder, wenn man Verweise zwischen Knoten ermöglicht, zu viele Knoten als „Bedeutung“ eines einzelnen Knotens erhalten kann.

Die nächste Theorie, die erwähnt wird, ist die Schema-Theorie, mithilfe derer Ereignisse und Abläufe beschrieben werden. Die bekanntesten Ansätze hierzu sind Scripts und Frames, die gleich erläutert werden. Wenn man ein passendes Schema gefunden

hat, muß man Variablen o. ä. belegen, mögliche Folgen inferieren und evtl. Erwartungen erzeugen. Auch hierbei gibt es Vererbungsmöglichkeiten.

Bei Frames gibt es zwei Arten von Objekten: Slots und Filler. Slots sind die Eigenschaften bzw. Parameter, die ein Frame haben kann, Filler die aktuellen Belegungen dieser Slots. Dabei gibt es Default-Belegungen und Methoden, die beim Belegen von Slots ausgeführt werden: *if-needed*, falls ein Slot interessant oder notwendig ist, *if-added*, falls ein Slot tatsächlich ausgefüllt wird. Frames sind also ein Konzept, das sich mit Objekten der Welt befaßt.

Im Gegensatz zu Frames sind Scripts ereignisorientiert. Wichtig sind dabei die beteiligten Rollen bzw. Gegenstände, die Eingangsbedingungen und Ergebnisse sowie der Ereignisablauf. Mithilfe von Tracks kann man ein Script je nach Situation und Belegung in verschiedene Abläufe aufteilen. Schlüsselbegriffe weisen auf das richtige Script hin, dies können z. B. Scriptnamen, Tracks oder Rollen sein. Nach einem kurzen Beispiel erwähnt der Autor Erweiterungen der Script-Theorie mit Plänen und Zielen.

Die Bedeutung eines Wortes innerhalb der Schema-Theorien ist mit dem Schema und den dadurch aktivierten Schemata gegeben. Habel erläutert nun kurz, daß man die vorgestellten Systeme zum einen deklarativ sehen kann, als Abart der Prädikatenlogik. Aber auch eine prozedurale Sichtweise ist möglich, z. B. wegen der *if-added* und *if-needed*-Mechanismen bei Frames. Dann erläutert er die Unterschiede bezüglich Vollständigkeit und Konsistenz bei deklarativem und prozeduralem Vorgehen. Zur Konsistenz erwähnt er kurz die Defaultlogik, die nicht-monotones Schließen benutzt und mit ihren Default-Regeln auch ein interessanter Ansatz ist.

Im vorletzten Teil geht Habel noch kurz auf offene Probleme ein: so ist manchmal bildhaftes Wissen vonnöten, oder das implizite Wissen um die notwendige und hinreichende Inferenztiefe abhängig vom Interesse bzw. der Relevanz des Schlusses. Auch erwähnt er, daß aus der Linguistik oder der Psychologie interessante Ansätze kommen.

Schließlich betont der Autor in der Zusammenfassung noch einmal, daß bei der Verarbeitung natürlicher Sprache das Lexikon sowohl deklarativ als auch prozedural sein sollte, und auch beim Abruf von Einträgen gewisse Prozesse starten könne sollte. Die Trennung von verschiedenen Wissenskategorien ist nicht praktikabel, da oft mehrere Arten von Wissen gebraucht werden. Schließlich betont er die Verknüpfung der Lexikographie mit der Wissensrepräsentation.

3 Computerlinguistik und das Lexikon

In ihrer Einleitung gehen Boguraev und Briscoe auf die Schwierigkeiten bei der Konstruktion eines Lexikons ein. Da es noch keine wohlformulierte Theorie über dessen Inhalt gibt und weil man es mit einer so großen Menge an Wörtern zu tun hat, gab es in den Achtzigern noch wenige adäquate maschinenlesbare Wörterbücher (MRDs, machine-readable dictionaries). Bei den existenten MRDs sehen sie die Vorteile darin, daß, da diese meist von gedruckten Wörterbüchern abstammen, sie von der Tradition der Druckwerke profitieren, und auch meistens eine große Menge Wörter abdecken, was den Computerlinguisten Arbeit erspart. Andererseits sind die gedruckten Wörterbücher für Menschen geschrieben, und setzen voraus, daß der Benutzer (englische) Wortdefinitionen verstehen kann. Deshalb dreht sich bei der Forschung viel um die Probleme, Informationen aus existenten MRDs in maschinenverarbeitbarer Form zu extrahieren und sie in bestehende sprachverarbeitende Systeme einzubauen. Diese Forschung nennen die Autoren „Computerlexikographie“.

Zur Verarbeitung natürlicher Sprache braucht man eine funktionierende Theorie. Auch wenn noch keine solche allumfassende Theorie formuliert wurde, konnte man adäquate sprachverarbeitende Systeme mit dem damaligen Forschungsstand bauen. Die meisten dieser Systeme arbeiten wissensbasiert, d. h. das nötige Wissen ist explizit eingebaut. Zu diesem Wissen zählen phonologische, morphologische, syntaktische, semantische und pragmatische Regeln, wobei letztere wenig mit dem Lexikon zu tun haben. Anhand eines Beispiels erläutern Boguraev und Briscoe jetzt einige solcher Regeln und führen nebenbei Phrasenstrukturgrammatiken ein. Das verwendete Lexikon zeigt, wie gewisse Informationen wie Features und Wortarten mit den formulierten Regeln zusammenspielen. Danach zeigen die Autoren, daß Bedeutung z. B. im Zusammenhang mit Präfixen ableitbar ist, aber daß es auch hierfür Ausnahmen gibt (think und rethink vs. produce und reproduce). Ein ideales Lexikon würde solche Regeln und Ausnahmen beinhalten. Viele Wörterbücher listen allerdings durch Regeln ableitbare Einträge explizit auf.

Im nächsten Abschnitt erwähnen die Autoren, daß die meisten sprachverarbeitenden Systeme kleine Lexika haben - meist, weil es zuviel Mühe ist, ein großes selber aufzubauen oder weil die Systeme nur Prototypen sind. Deshalb interessieren sich viele Forscher für MRDs, weil die Lexika ja gewissermaßen schon fertig sind. Aber weil die meisten Systeme unterschiedliche Formate für ihre Lexikoneinträge haben und auch verschiedene Ansprüche, ist die Frage, ob es machbar ist, eine einzige lexikalische Datenbank zu bauen, aus der die einzelne Systeme ihre Wörterbücher entnehmen. Eine solche Datenbank kann man nicht von Hand bauen, aber auch wenn man dafür elektronische Versionen von Wörterbüchern nimmt, hat man ein Problem mit der Informalität der Einträge.

Ein Lexikoneintrag besteht typischerweise aus dem Schlüsselwort mit Informationen über Aussprache und Schreibung, vielleicht auch Verwendung. In der Funktion folgt eine Beschreibung des Verhaltens, und dann die Bedeutung. In vielen MRDs sind mithilfe besonderer Notation gewisse Angaben notiert, z. B. bei Verben die Eigenschaften der Objekte, oder bei Nomen die Kategorie. Meist benötigt aber ein sprachverarbeitendes System viele dieser Informationen nicht. Auch ist die Beschreibung der Bedeutung des Wortes selber sprachlich gegeben, was bei sprachverarbeitenden Systemen eine Zirkel erzeugt - woher soll das System das Wissen um die Wörter nehmen, mit denen die Einträge definiert sind, wenn nicht aus dem Lexikon? Es stellt sich auch die Frage, ob man wirklich ein MRD in einem System benutzen kann, in dem beim Parsen Funktionen gestartet werden. Ein anderes Problem ist die Inkonsistenz bestimmter formaler Einträge - wenn kein formales System existiert, um diese Einträge eindeutig zu definieren, hängt das Ergebnis vom Ermessen der Lexikographen ab und ist dann maschinell fast unbrauchbar. Auch können Satz-Informationen in den Daten enthalten sein, die vollkommen unwichtig für ein sprachverarbeitendes System sind. Zirkuläre Definitionen stellen ein anderes Problem dar. Manchmal ist selbst die alphabetische Ordnung ein Problem, wenn z. B. Spracherkennung mit dem MRD erfolgen soll; dann wäre eine phonologische Ordnung vorzuziehen.

In einem Überblick über die Arbeit mit MRDs werden kurz Wortlisten zur Rechtschreibprüfung, semantische Taxonomien, die aus MRDs erstellt wurden, „Browsing“ oder das Auffinden von ähnlichen Wörtern aus verschiedenen Wörterbüchern, Sprachverarbeitung (Erkennung und Erzeugung), Parsing und semantische Verarbeitung als Anwendungsbeispiele von MRDs erwähnt. Bei der semantischen Verarbeitung wird das MRD als Wissensbasis verwendet. Auch Netzwerke, die man aus MRDs erstellen kann, werden angesprochen. Bei der Disambiguierung ohne syntaktische Verarbeitung haben sich MRDs auch als nützlich erwiesen. Da Wörterbücher meist schon hierarchisch organisiert sind, kann man daraus Ontologien erstellen, die oft verwendet werden. Einige Wörterbücher, die ein beschränktes Vokabular für ihre Definitionen wählen, können auch als semantische Datenbanken dienen. Nur in der Textgenerierung machen MRDs wenig Sinn, bis man mehr über die Prozesse bei der Wortauswahl weiß.

Die Anwendbarkeit von MRDs ist eine wichtige Frage. In diesem Zusammenhang muß man wissen, wie ausführlich ein Wörterbuch ist, wie es organisiert ist und wie man die Informationen extrahieren kann. Man sollte sich auch Gedanken darüber machen, was zu tun ist, wenn man die Grenzen des Wörterbuchs erreicht hat, da kein Wörterbuch je als vollständig bezeichnet werden kann. Probleme gibt es bei der Verlässlichkeit von MRDs. Da diese von Menschen gemacht wurden, schleichen sich Fehler ein, syntaktischer wie semantischer Art - vergessene Klammern, inkonsistente Formate, zirkuläre Definitionen, ungenaue oder redundante Angaben, usw. .

4 Vergleich der beiden Texte

Beide Texte stammen aus den Achtzigern, daher denke ich, daß sie vom Stand der Technik und Forschung durchaus vergleichbar sind. Im Ansatz unterscheiden sie sich allerdings, zum einen schon, weil Habels Text ein Artikel in einem Buch zur Lexikologie ist, der Text von Boguraev und Briscoe jedoch ein Einführungstext in einem Buch über Computerlexikographie. Zum anderen unterscheiden sie sich auch wie folgt:

Habel beschäftigt sich mit möglichen Formen von Lexika und stellt dabei verschiedene Möglichkeiten wie semantische Netze, Scripts oder Frames vor, welche meistens typische Werkzeuge der KI sind. Boguraev und Briscoe dagegen behandeln schon vorhandene Lexika, die eher die Form eines traditionellen Wörterbuchs haben, erwähnen aber auch Projekte, die aus MRDs semantische Netzwerke extrahieren können.

Dabei kümmert sich Habel nicht um die Machbarkeit von z. B. einer Ontologie von Frames: wer möchte schon von Hand alle Tätigkeiten des Lebens in einer Frame-Datenbank formulieren? Auch wenn diese Schemata genau formalisiert sind, sind die meisten seiner Ansätze nicht praktikabel. Zudem vermischt er die Semantik oft mit der Syntax - ein Lexikon muß nicht unbedingt die genauen Bedeutungen der Wörter formulieren, die Frage ist außerdem, wie das geschehen soll. Die vorgestellten Skizzen z. B. zu Schanks CDT mögen zwar recht eindrucksvoll sein, aber wie genau soll man solche Skizzen mit einem Computer verarbeiten, und woher bekommt man überhaupt solche Skizzen?

Boguraev und Briscoe sind da pragmatischer, wobei sie auch viel über tatsächliche Projekte berichten. Sie kümmern sich eher um die Probleme, wie man die Daten aus dem MRD verarbeiten kann und in welcher Form das geschehen sollte. Es ist keine dumme Idee, sich die Arbeit zu ersparen, ein maschinenlesbares Lexikon von Hand zu erstellen und auf schon vorhandenes zurückzugreifen, auch wenn sich dabei Probleme stellen wie die Verlässlichkeit des Lexikons. Heute erscheinen einem die erwähnten Lexika winzig klein, und es gibt auch schon Projekte, sich automatisch Korpora zu erstellen - damals war man noch nicht so weit. Das größte Problem ist sicherlich, wie man mit den Definitionen der MRDs umgehen soll, und der Zirkel, der bei sprachverarbeitenden Systemen dabei entstehen kann.

Beide Texte erwähnen semantische Netzwerke - Habel als wichtiges Element der Wissensrepräsentation, Boguraev und Briscoe im Zusammenhang mit maschinell aus MRDs erstellten Taxonomien. Dies scheint also zum einen eine KI-technisch begründete Technik zu sein, zum anderen auch u. U. computergestützt erstell- und benutzbar - wie man heute u. a. an Wordnet ¹ sieht, daß zwar nicht maschinell erstellt wurde,

¹<ftp://ftp.cogsci.princeton.edu/pub/wordnet>

aber häufig in Projekten, die eine semantische Datenbank brauchen, verwendet wird. Ich frage mich jedoch, wie man denn vorgehen würde, wenn man aus einem MRD ein semantisches Netzwerk erstellen wollte - die Hierarchie in Wörterbüchern ist nach meiner Erfahrung wenn vorhanden, dann eher vage, und zum Teil auch zyklisch - weshalb ein solches Vorhaben bestimmt nicht einfach ist.

Heute gibt es schon viele Wörterbücher in maschinenlesbarer Form, ob diese jedoch immer auch als MRDs verwendbar sind, ob aus Aufbereitungs- oder Lizenzgründen, ist die Frage. Aber auch die Korrektheit der MRDs dürfte sich gebessert haben, zum einen mit der Computerisierung der Verlage und auch durch technisch mögliche Korrekturüberprüfungen. Die Vision von einem einzigen Lexikon, aus dem verschiedene Projekte die für sie wichtigen Daten extrahieren, die Boguraev und Briscoe ansprechen, ist mit modernen Mitteln erreichbar geworden, aber wegen der gestiegenen Speicher- und Rechenkapazität vielleicht auch nicht mehr nötig.

Insgesamt sind beide Texte heute ein wenig überholt, aber das Fazit Habels, daß Wissensrepräsentationstechniken für Lexikographen wichtig sind, (zumindest wenn sie MRDs erstellen), und eine mögliche Schlußfolgerung Boguraevs und Briscoes, daß man sich nämlich beim Erstellen von MRDs auf Wörterbücher beziehen kann, gelten heute noch, und mit heutigen Programmiersprachen sind sie noch nicht einmal widersprüchlich.

5 Fazit

Beide Texte befassen sich mehr mit Lexika im allgemeinen und weniger mit ihren Einheiten. Wenn sie darauf eingehen, dann meinen beide meist Begriffsrepräsentationen und nicht Phrasen oder Wörter eines Satzes. Da aber beide nicht auf Phrasen eingehen, vernachlässigen sie Kollokationen jedoch.

Bei der Organisation eines Lexikons sollte man sich schon Gedanken über Phrasen machen, um möglichst effizient und natürlich mit Kollokationen umgehen zu können. In einem Lexikon im Sinne Habels ist dies jedoch weniger wichtig. Der andere Text erwähnt allerdings Phrasen nur am Rande, obwohl er sich mit Wörterbüchern befaßt, die mit Sicherheit auch Mehrworteinträge beinhalten.

Die große Frage ist jedoch, wie man Kollokationen in einem semantischen Netzwerk behandeln will. Dabei stellen sich Probleme wie die Einordnung, Repräsentation und Bedeutung solcher Phrasen. Wie wir im Laufe des Seminars festgestellt haben, sind Kollokationen auch in den aktuellen Wörterbüchern meist ungenügend abgedeckt und nicht reproduzierbar eingeordnet. Wenn man also dem Ansatz Boguraevs und Briscoes folgen will, müsste man zunächst ein Wörterbuch finden, daß solche Phrasen berechenbar einordnet und definiert. Auch stellt sich die Frage nach dem Format, wenn z. B. Adjektive erlaubt sind oder nicht, und welche. Das Definitionsproblem ist wie bei den anderen Wörtern nicht einfach, wobei Kollokationen auch Einfluß auf Stil und Intention haben, wenn sie z. B. ironisch gemeint sind.

Aber wie im Format der MRDs sich hoffentlich etwas verbessert hat, kann man auch mit dem Fortschreiten der Forschung zu Kollokationen und der Verbreitung der Ergebnisse hoffen, daß sich auch diese Situation zum Besseren wendet, auch wenn kein Wörterbuch je alle Wörter und Kollokationen auflisten kann.

A Literaturverzeichnis

HABEL, CHRISTOPHER „Das Lexikon in der Forschung der Künstlichen Intelligenz“ *Schwarze, Ch. & D. Wunderlich (Hrsg.), Handbuch der Lexikologie*, Königstein/Ts., 1985

BOGURAEV, BRAN & TED BRISCOE „Introduction“ *Boguraev, Bran & Ted Briscoe (Hrsg.), Computational Lexicography for Natural Language Processing*, New York: Longman, 1989