



Der neue Brockhaus: Einsatz von Sprachtechnologie und Wissensnetz

Bernd Kreissig, Mannheim

Menschliches Wissen und menschliche Sprache sind aufeinander bezogen und voneinander abhängig. Dies gilt auch und gerade bei den Bemühungen, Sprache und Wissen computergestützt so zu verarbeiten, dass für Menschen nutzbare maschinelle Wissensverarbeitung stattfindet. Jenseits von eher akademischen „Laborversuchen“ hat wissenschaftliche Forschungsarbeit auf diesem Gebiet die Verfügbarkeit umfangreicher, gepflegter und sauber codierter Wissensbasen sowohl der verwendeten Sprache als auch des abgebildeten Gegenstands zur Voraussetzung. In einem Forschungsprojekt der Firma Brockhaus Duden Neue Medien GmbH gemeinsam mit namhaften wissenschaftlichen Partnern, u. a. dem Fraunhofer-IPSI, dem IICM der TU Graz, dem IAI der Universität Saarbrücken und dem KNOW-Center, Graz, wurde diese Herausforderung erfolgreich angegangen. Für eine neue digitale Brockhaus-Auflage wurden Informationsrepräsentations- und Retrieval-Techniken entwickelt, die nicht auf domänenspezifischen Beschränkungen beruhend den gesamten enzyklopädischen Wissenskreis abdecken. Die entwickelten Formalismen erwiesen sich nicht nur zur Abfrage des enzyklopädischen Wissens in natürlicher Sprache als geeignet, sondern konnten auch darauf basierend neuartige Formen der Visualisierung von Wissenszusammenhängen implementiert werden. Die Ergebnisse zeigen zum einen Möglichkeiten erfolgreichen semantischen Retrievals ohne extensiven Einsatz formaler Metacodierungen, zum anderen weisen sie den Weg und die nächsten Schritte bei der Entwicklung noch leistungsfähigerer Mensch-Maschine-Schnittstellen.

Language technology and semantic networks for the new digital Brockhaus edition

Human knowledge and language refer to and depend on each other. This fact is highly relevant if knowledge and language should be automatically processed in a way that allows the engineering of human knowledge. Outside of academic laboratory scenarios, scientific research in knowledge and language processing needs large, well-kept and properly encoded knowledge bases of language and domain knowledge. This challenge was successfully addressed by the company Brockhaus Duden Neue Medien GmbH in a research project together with the Fraunhofer IPSI institute in Darmstadt, the IICM institute of the Technical University Graz, the IAI institute of the University Saarbrücken and the KNOW-Center at Graz. For the new digital Brockhaus edition, special information representation and retrieval technologies were developed that are not restricted to one domain but cover the whole encyclopaedic knowledge space. The newly developed techniques were proven to be applicable to natural language search as well as to innovative knowledge visualisation. The results show – on the one hand – the opportunities for successful semantic retrieval without extensive application of formal meta-encoding. On the other hand, they point at the next steps to take in order to develop still more effective human-machine-interfaces.

Die Erforschung der maschinellen Verarbeitung von Wissen und Sprache hat einen langen Weg sowohl schon hinter sich als auch noch vor sich. Dabei sind die verschiedenen Phasen dieses Weges jeweils davon geprägt, dass einerseits die Voraussetzungen und technischen Möglichkeiten sich stark verändert haben und dass andererseits jeweils andere Anwendungsfälle „in Mode“ waren. In den frühen Jahren maschineller Sprach- und Wissensverarbeitung wählte man sich in der theoretischen Bewältigung der zugehörigen Fragen relativ weit, lediglich die begrenzte Leistungsfähigkeit der Hardware

schien einen breiten Einsatz entsprechender Technologien temporär zu verzögern. Eine Anwendungs-„Modethema“ war z. B. die Sprachsteuerung von technischen Geräten. Entsprechend optimistische Prognosen waren gängig – das Ende von Tastaturen als Standard-Eingabegeräte wurden z. B. seit den 70er-Jahren immer wieder proklamiert, zugunsten von Sprach- und anderen Steuerungen, und ist doch nicht in Sicht.

Als es mit der Verfügbarkeit von mehr Rechenleistung darum ging, die Versprechen einzulösen, stellte sich heraus, dass die Realisierung von sprach- und wissensver-

arbeitenden Systemen doch schwieriger war als antizipiert. Dies betrifft sowohl den akustik-bezogenen Teil von Sprache („speech technology“), auf den ich hier nicht weiter eingehen will, als auch den morphosyntaktischen und semantischen Aspekt („language technology“). Es stellte sich zunächst heraus, dass die primär regelbasierten Systeme bestenfalls für sachlich jeweils eng begrenzte Wissensmodellierungen geeignet waren – immerhin entstanden nun brauchbare Expertensysteme. Der nun immer prominenter werdende Ansatz, die Fragestellung mit primär statistischen, möglichst sprachunabhängigen (im Sinne von „sprachenübergreifenden“) Mitteln zu lösen, stellte prinzipbedingt noch höhere Anforderungen an die verfügbare Rechenleistung. Er stand und steht auch im Zusammenhang mit den neueren Anwendungen-„Modethemen“ rund um das Problem der Beherrschung der immer größeren Informationsmengen, insbesondere der digital verfügbaren Informationsmenge. Statistische Verfahren konnten aber naturgemäß die Frage maschineller Wissensverarbeitung ebenfalls nicht zufriedenstellend allgemeingültig lösen.

Sprach- und wissensverarbeitende Systeme haben sich meiner Einschätzung nach bislang auch deshalb nicht auf breiter Basis durchgesetzt, weil der Abstand zwischen den aufgebauten Erwartungen einerseits und dem real Möglichen andererseits einfach zu groß war. Hieraus ergibt sich aber nun auch eine Chance: Zum einen sind die Erwartungen nicht mehr so unrealistisch wie z. B. vor 20 Jahren, zum anderen ist die Leistungsfähigkeit der Systeme seitdem beträchtlich gestiegen. Vor allem aber ist es mittlerweile akzeptierte und gelebte Praxis, dass maschinelle Sprach- und Wissensverarbeitung Zusatznutzen schafft, aber (noch) nicht die primär bzw. gar ausschließlich eingesetzte Technik für die jeweilige Aufgabe darstellt. In einer solchen Konstellation ist kontinuierliche Steigerung und Verbesserung möglich, anstatt dass die Akzeptanz an den noch fehlenden zehn oder zwanzig Prozent gleich zu Beginn nachhaltig verloren geht. Ich werde diesen Punkt gleich noch beispielhaft demonstrieren.



PRAXISBEISPIEL BROCKHAUS

Die Beiträge dieses Schwerpunktheftes machen einmal mehr deutlich, dass der Aufbau von maschinell nutzbaren Wissensstrukturen an natürlicher Sprache nicht vorbeikommt. An die Vision, man könne das Weltwissen abstrakt, in sprachunabhängigen oder -übergreifenden Symbolsystemen ablegen und verarbeiten, glaubt heute niemand mehr ernstlich. Wer heute mit Rechnern Wissen verarbeitet, verwendet die Token, die unserer natürlichen Sprache angehören oder entstammen – zu eng ist die Art, wie wir wissen und denken mit der Art, wie wir sprechen, verbunden. Das bedeutet in Konsequenz, dass derjenige, der über hochwertig modelliertes und codiertes Sprachwissen verfügt, gleichzeitig über besonders gute Möglichkeiten des Aufbaus maschinell verarbeitbarer Wissensstrukturen verfügt. Hier liegt für einen Wörterbuchverlag wie Duden ein immenses Zukunftspotenzial. Umgekehrt gilt allerdings auch: Wer Sprache maschinell hochwertig codieren will, kommt um die semantische Ebene von Wörtern nicht herum – das Wissen über die Dinge dieser Welt und ihre Struktur bzw. Strukturierbarkeit ist hier unverzichtbar. Über solches formalisiertes Weltwissen verfügen wir als Lexikonverlag Brockhaus nun glücklicherweise ebenfalls, und das versetzt uns in eine nochmals verbesserte Ausgangsposition.

Ich möchte Ihnen im Folgenden nun darstellen, auf welche Weise wir begonnen haben, diese Potenziale zu nutzen. Unser Verlag hatte bereits 1992, also lange vor der Erfindung von XML und als erster großer Verlag in Deutschland damit begonnen, die Lexikon- und Wörterbuchinhalte in einem SGML-Redaktionssystem zu pflegen. Vor fünf Jahren haben wir dann gemeinsam mit dem Fraunhofer-IPSI sowie der Firma intelligent views, einer IPSI-Ausgründung und bis heute enger Partner von uns, ein Modell für eine Sprach-Wissens-Modellierung entwickelt und als Software implementiert¹. Das entwickelte Objektmodell geht von Bezeichnern (Lemmata) und Bezeichnetem (Konzepten) aus; ein „Term“ ist dann die Unifizierung aus beidem, d. h. ein Lemma in einer bestimmten Bedeutung. Das lässt sich etwa mit folgendem Schema veranschaulichen:



Abbildung 1: Schema der Lemma-Term-Konzept-Beziehungen

An den verschiedenen Objekttypen lassen sich nun zum einen unterschiedliche Eigenschaften verankern (z. B. ein Genus an einem Lemma-Objekt) – wichtig und nützlich, um z. B. klassische Wörterbücher produzieren zu können. Für die Entwicklung innovativer elektronischer Medien fast noch interessanter sind nun aber die Relationen zwischen Objekten. Und zwar nicht nur die eben genannten Unifizierungen, sondern gerade auch die Relationen zwischen gleichen Objekttypen, also z. B. die Relation zwischen zwei Konzepten, etwa dass eine Pflanze ein Lebewesen ist, usw.

Damit verfügten wir nun über ein Redaktionssystem, welches nicht nur die zur Wörterbuchproduktion nötigen Angaben aufzunehmen in der Lage war, sondern gleichzeitig auch darüber hinausgehende Wissensstrukturen. Es ging und geht seitdem also darum, diese weitergehenden Wissensstrukturen zu definieren und aufzubauen. Möglichkeiten hierzu gibt es potenziell unendlich viele. Ein besonderer Vorteil von Wissensnetz-Modellierungen gegenüber klassischem Markup liegt nun darin, dass man sich nicht vorab auf ein Content-Modell festlegen muss, sondern im Prinzip verschiedene Relationsnetze parallel nebeneinander existieren lassen kann. In der Praxis überlegt man sich u. a. wegen des damit verbundenen Pflegeaufwands natürlich zunächst eine möglichst schlanke, effiziente Modellierung.

Entscheidend für die Wahl des Modells und der Mittel zur Umsetzung solcher Relationsnetze ist, wenig überraschend, auch hier der intendierte Einsatzzweck. Der lag in unserem Fall vor uns im bevorstehenden Jubiläum des Brockhaus-Verlags. Zu diesem 200. Geburtstag sollte nämlich die berühmte Enzyklopädie in einer neuen, 21. Auflage erscheinen. Und sie sollte dies nicht nur, wie allgemein erwartet und vorausgesetzt, in Bezug auf Inhalt, Einband, Gestaltung, Satztechnik, Herstellung usw. in perfekter Form tun. Der Auftrag des Verlagsvorstands an die Brockhaus Duden Neue Medien GmbH lautete, das Werk gleichzeitig in den digitalen Offline- und Onlinemedien zu konzipieren und zu realisieren. Die digitale Ausgabe sollte um nichts weniger als die gedruckte Ausgabe in jeder Hinsicht Maßstäbe setzen. Ich berichte das deshalb so ausführlich, weil es hier über „Chancen des automatischen Aufbaus von Wissensstrukturen“ gehen soll. Nun, das war eine ganz handfeste Chance, nämlich eine Nachfrage, etwas zu tun, was es in der Form bislang noch nicht gab. Wir haben diese Aufgabe auf verschiedenen

Ebenen gelöst. So haben wir z. B. die Software so implementiert, dass wir sie auf einem USB-Stick ausliefern konnten, der ohne Installation lauffähig ist und mithin den Inhalt der 30 Bände maximal transportabel macht. Der eigentliche Clou liegt aber in der Software selbst, wo wir Wissensstrukturen in einen Mehrwert bei der Wissenserschließung umgesetzt haben. Es ging uns dabei darum, neben dem traditionellen, klassischen Nachschlagen eines Stichworts neue, explorative Möglichkeiten des Wissenszugangs zu schaffen. Die prominentesten beiden dieser Möglichkeiten möchte ich nachfolgend beschreiben. Die erste nennen wir „3-D-Wissensraum“. Dabei geht es darum, lexikalische Wissenseinheiten, also Artikel, in Beziehung zu anderen zu bringen, und vor allem dieses zu inszenieren. Das sieht im Ergebnis so aus wie in Abb. 2 gezeigt.

Im Prinzip haben wir es hier mit zwei Arten von automatischem Aufbau von Wissensstrukturen zu tun, die in die unterschiedlichen räumlichen Dimensionen projiziert sind. Horizontal wurde eine assoziative Bezugsstruktur errechnet, letztlich also nichts anderes als ein Dokumentcluster von Lexikonartikeln, bei denen verschiedene Merkmale bzgl. ihrer Relevanz zueinander gemessen werden. Vergleichsweise einfache Gestaltungsmittel sorgen dann aber für ein sehr lebendiges Bild: die verwandten Artikel sind thematisch sortiert, d. h. die verschiedenen Farben stehen für unterschiedliche Sachgebiete. Die verschiedenen Eintragstypen werden durch unterschiedliche Figurenformen wiedergegeben, z. B. ein Artikel über eine Person mit einer Spielfigur, ein Sachartikel als kleine Tonne. Die unterschiedliche Größe der Figuren zeigt den Umfang des dahinterstehenden Artikels an, der durch Mausclick aufgerufen werden kann. Orthogonal zu diesen assoziativen Verbindungen, sozusagen in der Vertikalen, sind an vielen Artikelfiguren sachsystematische Verbindungen geknüpft. Im vorliegenden Beispiel öffnet ein Rechtsklick auf die Figur „Hellpach“ ein Kontextmenü, welches die Person als deutschen Psychologen ausweist und gleichzeitig die Verbindung zu weiteren Einträgen des gleichen Typs anbietet. Auch diese Wissensstruktur ist eine zunächst automatisch aufgebaute; auch hier hat uns wieder das Fraunhofer-IPSI unterstützt, und wir haben eine sechsstellige Anzahl von solchen Relationen aus unseren eigenen Lexikon- und Wörterbuchdefinitionen herausgearbeitet² und dann die Ergebnisse selbstverständlich noch einmal durch unsere Redaktion qualitätsgesichert. Das Schöne an dieser Nutzung automatisch aufgebaute Wissensstrukturen ist, dass sie – übrigens relativ begeistert begrüßt – Nutzen schaffen, lange bevor sie perfekt funktionieren und/oder vollständig sind. Wenn in einem automatisch

1 Vgl. www.ipst.fraunhofer.de/~rostek/alexa-etali-irec2002.pdf
 2 Vgl. dazu den Beitrag von A. Dirsch-Weigand und I. Schmidt auf S. 367.



PRAXISBEISPIEL BROCKHAUS

Insgesamt zeichnet sich das System durch eine hohe Robustheit aus; damit geht einher, auch nicht optimale Ergebnisse im Suchergebnisraum zu haben. Der Fall, dass zu einer Frage gut passende Einträge überhaupt nicht gefunden werden, ist deutlich seltener. Für mich ergibt sich aus dieser Konstellation die nächste Chance des automatischen Aufbaus von Wissensstrukturen: Wenn man schon so weit ist, aus einem großen Repository eine überschaubare Menge an zu einer Frage gut passenden Dokumenten zu selektieren, dann müsste die Extraktion der konkret benötigten Angabe daraus – also eine echte „Antwort“ auf die Frage – doch auch in Reichweite sein? Das künftig zu erarbeiten ist eine der Herausforderungen und Chancen, denen sich die Brockhaus Duden Neue Medien GmbH in der Zukunft stellen wird.

Nachschlagewerk, Computerlinguistik, semantisches Netz, Wissensrepräsentation, Information Retrieval, Informationssystem

DER AUTOR

Bernd Kreissig



(43) studierte in Oberursel/Ts., Jerusalem, Ft. Wayne (USA) und Heidelberg evangelische Theologie. 1994 Eintritt in den Verlag Bibliographisches Institut & F. A. Brockhaus AG als Anwendungsprogrammierer; seit 2001 Geschäftsführer der Brockhaus Duden Neue Medien GmbH.

Brockhaus Duden Neue Medien GmbH
Dudenstr. 6, 68167 Mannheim
kreissig@bifab.de