Kurt Eberle

Textverstehen und Wissensmanagement

Der Erfolg des Internet und der Anwendung firmeninterner Intra-Netze bedeutet, dass immer mehr Wissen elektronisch verfügbar wird. Dasjenige Unternehmen das schnell das zu einem bestimmten Thema relevante Wissen finden, aufbereiten, zusammenstellen, seinen Kunden in deren Sprache präsentieren kann hat in der globalen Wirtschaftswelt einen gewichtigen Vorteil seinen Konkurrenten gegenüber. Es kann dies umso schneller je mehr Werkzeuge zum Wissensmanagement, und das bedeutet zuallererst zur automatischen Sprachverarbeitung, es einsetzen kann. Damit ist diese Verarbeitung, das *Rechnen mit Sprache*, ein Markt und vor allem ein Zukunftsmarkt. Dieser Zusammenhang begründet die wachsende Wahrnehmung und Anerkennung der Computerlinguistik im großen Publikum und damit auch den beginnenden kommerziellen Erfolg, sichtbar an der wachsenden Anzahl von Firmengründungen zu diesen Themen in der jüngsten Vergangenheit.

1 Rechnen mit Sprache

Rechnen mit Sprache ist zunächst, wie das Rechnen mit Zahlen, die Erzeugung und der Vergleich korrekter Formen und Ausdrücke durch Manipulation von Zeichen auf der Basis geeigneter Produktionssysteme, so die Einsicht und der Beitrag der theoretischen Informatik bzw. allgemeinen Computerwissenschaften.

In den frühen Jahren der Computerlinguistik bedeutete diese Perspektive, die (vorwiegend aus der strukturalistischen und transformationellen Forschung) vorliegenden empirischen Ergebnisse zur morphologischen Merkmalanalyse des Basisinventars der Sprachen und zur Beschreibung des syntagmatischen Verhaltens durch Ersetzungsregeln einer formalen Grammatik als Programme zu formulieren. Verstanden als Theorie, beschrieben solche Programme formale Sprachkompetenz. Praktisch bedeutete die Möglichkeit, Korrektheit von Wortformen und Phrasen vorhersagen, bzw. solche korrekten Ausdrücke ableiten zu können, die Verfügbarkeit recht funktionstüchtiger so genannter Spell- und Grammar-Checker, wie sie heute Bestandteil jedes besseren Textverarbeitungssystems sind, und anderer hilfreicher Werkzeuge für die formale Analyse und Datenerfassung zu und aus Sprachmaterial. Über eine abstrakte syntaktische Repräsentationsebene zu verfügen, bedeutete auch, Übersetzungssysteme herstellen zu können, die besser waren als einfache Wort-zu-Wort-Übersetzungen. Das waren frühe Erfolge, die im Wesentlichen in die 60er-Jahre zurückreichen.

Allerdings war immer schon klar, dass spannendere (und kommerziell lukrativere) Aufgabenstellungen, wie das automatische Beantworten inhaltlicher Fragen zu Texten

(das sog. *information retrieval*) oder das automatische Zusammenfassen von Texten (*summarization*) oder die qualitativ gute, bedeutungserhaltende maschinelle Übersetzung, die kein ausführliches Postediting mehr verlangt, nicht nur die Form, sondern die Bedeutung der Zeichen und Ausdrücke und ihrer Kontexte miteinbeziehen müssen. Im Übrigen auch einige simplere computerlinguistische Anwendungen, dann wenn große Genauigkeit wünschenswert ist. Ein aktuelles Beispiel ist die Orthografiekonversion. Weil die Entscheidung bei einigen Konversionstypen ohne gutes Verständnis des Kontexts nicht automatisch getroffen werden kann, bietet leider kein auf dem Markt befindliches Produkt die äußerst wünschenswerte Konversion ohne interaktive Korrektur ausschließlich im Batch-Modus an.

Der Wunsch, in dieser anspruchsvollen Weise mit Sprache rechnen zu können, ist sehr alt, belegt seit dem hohen Mittelalter mindestens.

Nachdem Montague in den frühen 70er-Jahren einen Weg aufzeigen konnte wie natürliche Sprachen als spezielle formale Sprachen betrachtet werden konnten und damit es erlaubte, den Apparat der mathematischen Logik wie er seit Frege entwickelt worden und für die Computerwissenschaften als solche konstitutiv war für die automatische Sprachverarbeitung anzuwenden, begann eine Phase nahezu euphorischen Forschens, geprägt von der optimistischen Erwartung, dass durch die Abbildung von Texten auf Ausdrücke einer inhaltlich präzisen Bedeutungssprache mit regulären logischen Gesetzmäßigkeiten die algorithmische Modellierung von Textverstehen und, allgemeiner, des menschlichen Denkens greifbar nahe schien.

2 Gordischer Knoten Textverstehen

Die Hoffnung, dass die Qualität computerlinguistischer Anwendungen durch semantische Verfahren auf ein neues, bisher nicht gekanntes Niveau zu heben sei, wurde in zahlreichen Forschungsprojekten zu Frage-Antwortsystemen, zu maschineller Übersetzung, allgemein zu automatischem Wissensmanagement bis heute nicht wirklich bestätigt, zumindest nicht für Betrachter aus dem industriellen Umfeld, deren Urteil gerade unter dem Aspekt der gewachsenen kommerziellen Relevanz der Thematik immer wichtiger wird.

Schon Anfang der letzten Dekade wurde am Beispiel der Maschinellen Übersetzung auf der lange nachwirkenden TMI-92-Konferenz in Montréal hochkarätig über das Für und Wider der in der beschriebenen Weise analytischen Sprachverarbeitung gestritten und diskutiert, ob es nicht besser sei, die Analyse, vor allem die semantische, durch statistische Betrachtungen zu ersetzen. Das Hauptargument gegen den praktischen Nutzen semantischer Verfahren in der Computerlinguistik war, und ist seither geblieben, die schiere Größe des für die Interpretation notwendigen semantischen Hintergrundwissens und die Ambiguität der natürlichen Sprache, die ja Frege schon als einer

formalsemantischen Betrachtung zuwiderlaufend beklagt, die sich nicht nur in der Mehrdeutigkeit von Wörtern, sondern auch in strukturellen Mehrdeutigkeiten, in der vom Zweck abhängenden Bedeutung von Redebeiträgen zeigt und damit in der Folge auch in der notwendigen Schwäche von Beweisverfahren für Texte, deren Anwendung gleichwohl äußerst aufwändig ist.

Andererseits, so mittlerweile die Erkenntnis, haben statistische Methoden zwar ihre Vorzüge, sind aber, vor allem dort, wo es um neue Information, um die genaue Wiedergabe von Information, um selten durchgeführte Schlüsse geht, sprich überall dort, wo wegen der geringen Häufigkeit des Vorkommens die statistische Basis fehlt, prinzipiell im Nachteil.

Wenn Semantik also unabdingbar scheint, aber auf Grund der Mehrdeutigkeiten und der Informationsmenge schwierig angemessen operationalisierbar, wie ist dann zu verfahren?

3 Quo vadas, Computerlinguistik?

Neuere Entwicklungen im WWW, wie die Initiative zum *semantic web*, weisen den Weg. Es geht um sparsame Verwendung von semantischer Information: Also durchaus semantische Repräsentationen von Texten um einen hinreichenden Abstraktionsgrad für Aufgaben wie Klassifikation, Suche und Übersetzung zur Verfügung zu haben, basierend auf der semantischen Kennzeichnung der Wörter und deren Selektionsbeschränkungen vor dem Hintergrund semantischer Hierarchien, aber tendentiell kein Einsatz von Theorembeweisern über expressiven Sprachen.

Viel versprechend hinsichtlich des durch die Mehrdeutigkeiten gegebenen Komplexitätsproblems der Lesartenverwaltung ist, die Verzahnung der Beiträge der verschiedenen Analyseniveaus weiter voranzutreiben und, vor allem, früh in der Analyse zur Wirkung zu bringen, um eine hohe und schnelle Filterwirkung und Minimierung der Lesarten zu bewirken. Weil aber häufig Mehrdeutigkeiten verbleiben werden (und für bestimmte Zwecke, wie die bedeutungserhaltende Übersetzung, auch dürfen), ist die Weiterentwicklung von Theorien zur Repräsention und zum Umgang mit unterspezifizierter Information wichtig. Dort wo jedoch eine spezifische Lesart gefordert und mit den gegebenen Mitteln nicht ableitbar ist, müssen die Analysen gewichtet werden und die im gegebenen Kontext vernünftigste Rangordnung hat sicherlich mit Erfahrungswerten und demzufolge mit Statistik zu tun. Verzahnung von Komponenten wird also auch bedeuten müssen, statistische Verfahren in die Analyseprozesse zu integrieren. Die Frage ist also nicht und wird nicht sein, ob semantische Analyse oder statistisches Modell, sondern welche Gestalt die geeignete Integration der Verfahrenstypen hat.

Weil sich die Ergebnisse der Computerlinguistik immer stärker an den Erwartungen und Kriterien von *Kunden*, nicht von fachlichen Gutachtern, orientieren und bewähren

werden müssen, wird diese Roadmap flankiert werden müssen durch den Auf- und Ausbau einer ingenieurmäßigen Betrachtungsweise die wahrnimmt, dass für Anwender wichtig ist,

- dass die in der Praxis beobachtbaren Phänomene relativ vollständig abgedeckt werden, dass es im Falle eines Analysesystems also nicht genügt, gut mit Tempus umgehen zu können, aber nicht mit Pronomen,
- dass die Phänomene relativ gleichmäßig gut behandelt werden, also bei Übersetzungen beispielsweise, nicht eine exquisite Übersetzung neben einem völlig misslungenen Satz steht,
- dass in jedem Fall häufige Phänomene priorisiert vor weniger häufigen Phänomenen behandelt werden, dass ein System also erst dann seltene komplexe wh-Bewegungen hervorragend beherrschen braucht, wenn es auch mit den äußerst häufigen Phänomenen Ellipse und Koordination zufrieden stellend umgehen kann,
- dass Systeme robust, schnell, ergonomisch gut und dh. meist auch integriert in den allgemeinen Workflow sein sollen.

Dies bedeutet zuallererst, dass die Zeit der Spielsysteme, der Prototypen, die wenige Phänomene in Abstraktion aller anderen vor kleinem Fragment behandeln können vorbei sein muss. Wichtig ist, kumulativ und integrativ vorzugehen, um zu Systemen mit einigermaßen zufrieden stellender Abdeckung zu gelangen.

Die eigene langjährige Erfahrung mit der Weiterentwicklung eines großen kommerziellen Übersetzungssystems zeigt, dass ein großes Qualitätspotenzial im sauberen und geduldigen Zuendeformulieren von nicht notwendig im Rampenlicht stehenden Analysekomponenten liegt und es erklärt warum manche industrielle Entwicklung die auf relativ simplen Technologien aus der Frühphase der Computerlinguistik beruht, aufs Ganze gesehen, einfach auf Grund ihres Ausarbeitungsgrades, bessere Ergebnisse zeitigt, als so mancher unterdessen entstandener theoretisch tiefgründiger (z.T. auch tiefgründelnder) Forschungsprototyp. Es ist bezeichnend, was Tests deutlich machen, dass viele praktisch genutzten Systeme selbst geschlossene Bereiche, wie die Morphologie, nicht sauber und vollständig behandeln und es ist auch noch längst nicht jede (einfache) syntaktische Regel geschrieben.

Daneben muss Information aus verschiedenen Projekten zusammenfließen können, was bedeutet, dass die Bedeutung der Ausarbeitung von Standards für linguistische Beschreibungen wie OLIF oder SALT zunehmen wird und es muss Information aus vorhandenem Textmaterial möglichst direkt in das System einfließen, d.h. das System muss linguistisches Wissen möglichst automatisiert aus Corpora lernen können.

Mit der Menge der Information im System wächst die interne Komplexität. Die Erfahrung zeigt, dass die linguistische Extrapolation nicht linear zu haben ist, sodass der Umgang mit großen Datenmengen und die Laufzeitoptimierung von Algorithmen zu einem bedeutenden Thema wird.

Die Devise muss heissen linguistisches Software-Engineering, Standardisierung, Integration, Optimierung und statistisch fundiertes Lernen.