In: Harald H. Zimmermann (Hrsg., 1983): Informationswissenschaft an der Universität des Saarlandes. Bericht der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarlandes, 46-52

Harald H. Zimmermann

Maschinelle Verarbeitung natürlicher Sprache an der Universität des Saarlandes. Stand und Perspektiven aus informationswissenschaftlicher Sicht

1. Entwicklung

1.1 Grundlagen

Seit Anfang der 60-er Jahre gibt es an der Universität des Saarlandes Forschungen zur maschinellen Verarbeitung natürlicher Sprache. Neben dem Team des Sprachforschers Hans Eggers mit dem Schwerpunkt der Syntax des Deutschen war dies der Mathematiker Johannes Dörr, der mit seinen Mitarbeitern erste Ansätze zu formalen Grammatiken und entsprechenden Analyse-Algorithmen entwickelte; der "Ruhm", ein erstes Übersetzungsprogramm in Saarbrücken entwickelt zu haben, kommt jedoch dem Physiker Hubert Martin zu, der zusammen mit seiner Frau, einer Lateinlehrerin, auf der legendären "Zuse" (Z 22) ein System zur Übersetzung der Anfängerlektionen des "Fundamentum Latinum" vom Lateinischen ins Deutsche erstellte.

Nachdem sich vor allem die Aktivitäten der Eggers'schen Forschungsgruppe zunehmend von traditionell-statistischen Auswertungen zu maschinell-algorithmischen Untersuchungen gewandelt hatten (erstes wichtiges Resultat ist die Erstellung eines allgemeinen syntaxorientierten Analysesystems zum Deutschen (/1/) wurde Mitte der 60-er Jahre eine gemeinsame Forschungsgruppe Eggers/Dörr (Germanistik/Angewandte Mathematik) von der Deutschen Forschungsgemeinschaft mit der Bewertung eines Übersetzungssystems Russisch-Englisch und dessen Fortentwicklung zu einem System Russisch-Deutsch betraut. Grundlage war das System SYSTRAN (in seiner damaligen Form); hieraus entwickelten sich eigene Modelle und ein erstes eigenes Projekt zur Übersetzung Russisch-Deutsch. Am Anglistischen Institut der Universität entstand gleichzeitig ein Projekt zur Erstellung eines umfassenden maschinellen Wörterbuchs (Grundlage: Shorter Oxford English Dictionary), allerdings waren die Forschungsziele eher historisch-kritisch - im Gegensatz zu den sprachsynchronen Ansätzen in der Germanistik. Die Benutzung des gleichen Rechners (zunächst einer Philips Electrologica X1, später einer Control Data CDC 3300) brachten die Gruppen einander näher, Entwicklungen in der Romanistik zur Corpusanalyse (Hans-Ludwig Scheel) und zur Neologismenforschung im Französischen kamen hinzu, eine "Informatik" (Günter Hotz) begann sich aus der Angewandten Mathematik zu entwickeln - so dass zu Beginn der 70-er Jahre die Überlegungen konkrete Formen annahmen, die verschiedenen Ziele und Ansätze in koordinierte oder auch gemeinsame Forschungen einzubringen. Mit Unterstützung der Deutschen Forschungsgemeinschaft wurde 1972 ein Sonderforschungsbereich "Elektronische Sprachforschung (SFB 100) errichtet, der soeben (1983) unter seinem Sprecher Wolfram Wilß (der zugleich ein Projekt Automatische Englische Analyse einbrachte) in seine wohl letztmalige "Verlängerung" (bis 1986) eintritt.

Trotz der z.T. unterschiedlichen Aufgabenstellungen in der Ausgangsphase konnten die Forschungen sich zunehmend stärker auf erreichte Zwischenziele abstützen. Eine zentrale Rolle spielte dabei die Entwicklung allgemein verwendbarer Software.

Einerseits diente diesem Ziel die Konzeption einer speziellen Programmiersprache (COMSKEE) für die Zwecke der Sprachdatenverarbeitung. Eine entsprechende Realisierung in Laborform liegt inzwischen vor: sie wurde vom Teilprojekt Informatik (Hotz) in die Projektarbeiten eingebracht und wird inzwischen von allen übrigen Arbeitsgruppen testweise eingesetzt.

Eine weitere grundlegende Entwicklung stellt das sog. "Saarbrücker Übersetzungssystem" (SUSY) dar. Obwohl es aufgrund der technischen Gegebenheiten der 70-er Jahre eine relativ traditionelle Software-Basis hat, zeichnet es sich konzeptionell dadurch aus, dass es weitestgehend rechnerunabhängig programmiert ist (zur Zeit lauffähig auf einer TR 440 und einer SIEMENS 7561) und in weiten Teilen auch unabhängig ist von den Ausgangs- oder Zielsprachen, die verarbeitet werden sollen. Insoweit - sieht man von kleineren Teilen einmal ab - werden die "Regeln" (und Lexika) zur maschinellen Bearbeitung der natürlichen Sprachen Deutsch, Englisch, Französisch und Russisch, die gegenwärtig von den verschiedenen Teilprojekten erstellt und getestet werden, durch ein allgemeines Instrument unterstützt.

1.2 Anwendungen

Als Mitte der 70-er Jahre mit SUSY ein erstes Simulations- und Modellsystem (noch in rudimentärer Form) "technisch" vorlag, konnte daran gedacht werden, nicht nur sprach(system)spezifische Forschungen zu verfolgen (wie es im wesentlichen Aufgabe der Projektgruppe des SFB ist), sondern Fragen der Anwendbarkeit bzw. Anwendung anzugehen. Dies geschah - unter Verwendung von Teilen des SUSY-Systems, insbesondere der deutschen Sprachanalyse - zunächst an der Universität Regensburg. Als Anwendungsbereich wurde die Fachinformation und hierbei das Problemfeld Texterschließung (sog. maschinelle Indexierung) ausgesucht. Angesichts der ungeheuren Publikationsflut erschien es sinnvoll, zu prüfen, ob bestehende Retrievalsysteme dadurch verbessert werden können, dass qualitativ höherwertige Verfahren zum Einsatz kommen.

Nachdem die Ergebnisse recht viel versprechend waren /2/, werden sie - nunmehr an der Universität des Saarlandes, Fachrichtung Informationswissenschaft - in größeren Tests mit sog. "Anwenderdaten" (z.B. Daten des Deutschen Patentamts) fortgesetzt.

Ein weiteres Desiderat in der Fachinformation ist natürlich die maschinelle bzw. die maschinengestützte Übersetzung. Bislang sind hierzu keine Verfahren "am Markt" verfügbar, die automatisch für beliebige Texte eine Übersetzungsleistung erbringen, die der eines üblichen menschlichen Übersetzers entspricht. Hierzu sind aber mindestens zwei Alternativen denkbar: Einerseits ist zu prüfen, ob es nicht möglich ist, vollautomatische Übersetzungen zu erreichen, die informativ und brauchbar sind. Sofern dieses Ziel inhaltlich erreicht und kostengünstig realisiert werden kann, ließe sich wenigstens ein Teil der Problematik der Sprachbarrieren reduzieren. (Angesichts der Bedeutung, die dieser Entwicklung im Grundsatz zukommt, ist es verwunderlich, dass nicht schon längst erheblich größere systematische Anstrengungen hierzu national wie international gemacht wurden. Ein neuerlicher - wenn auch relativ bescheidener - Ansatz kann in den Planungen und Entwicklungen des Europäischen Übersetzungssystems EUROTRA bei der Europäischen Gemeinschaft gesehen werden, an dessen Konzeption und Entwicklung Saarbrücker Forscher erheblichen Anteil haben.) Ein Projekt (SUSY-DJT) mit einem entsprechenden Forschungsziel, am Modell (auf der Grundlage der SUSY-Software) zu zeigen, dass und inwieweit eine Informativ-Übersetzung machbar und nützlich erscheint, wurde 1983 mit Unterstützung des BMFT in Angriff genommen /3/.

Eine weitere Alternative ist es, den menschlichen Übersetzer durch Computer bei seiner Arbeit so zu unterstützen, dass damit zu ökonomischeren Bedingungen hoch qualifizierte Übersetzungen möglich werden. In diese Richtung gehen international eine Reihe von Forschungen (z.B. bei der EG die SYSTRAN-Anwendung); MDT-Systeme und Microcomputer werden zunehmend mit derartigen Funktionen ausgestattet. Auch an der Universität des Saarlandes wird ein Forschungsvorhaben (wiederum unter Verwendung der SUSY-Basissoftware) mit Unterstützung des BMFT (SUSY-BSA) durchgeführt, bei dem untersucht werden soll, in welcher Form ein Übersetzungssystem in den Arbeitsplatz eines Übersetzers (Modell "SUSY" beim Bundessprachenamt - BSA -) einzupassen ist.

2. Transferproblematik

Es soll im vorliegenden Zusammenhang nicht die Frage beantwortet werden, inwieweit nun all diese Forschungen funktional bereits einen realen Beitrag zur Erreichung eines speziellen Ziels der informationswissenschaftlichen Forschung, nämlich der Überwindung der Sprachbarrieren, gebracht haben. Dazu sind die Forschungs- und Rechenschaftsberichte der Projekte heranzuziehen bzw. abzuwarten. Es soll vielmehr an diesem Exemplum von Forschung und Entwicklung zu einem allgemeinen Ziel der Informationswissenschaft ein Beitrag gebracht werden, der zugleich von dem Autor "hautnah" erlebt und erfahren wurde (und wird) - dem Problem des Transfers von Wissen bzw. von Technologie und damit zusammenhängenden Rahmenbedingungen.

2.1 Komponenten des Transfers

Das grundsätzliche Modell des Technologie-Transfers ist in der Literatur genügend beschrieben /4/, als dass es hier nochmals zu begründen wäre. Seine Komponenten lassen sich auf die Situation der Forschung und Entwicklung zur maschinellen Sprachdatenverarbeitung an der Universität des Saarlandes wie folgt übertragen:

Grundlagenforschung:

Die Arbeiten des Sonderforschungsbereichs sind bei aller "empirischen" Orientierung (man vergleiche z.B. die theoretischen Arbeiten des Sonderforschungsbereichs 99 in Konstanz zur Sprachforschung) der Grundlagenforschung zuzuordnen. Im Wesentlichen stehen Untersuchungen zur natürlichen Sprache als System im Vordergrund, der Computer ist ein Instrument zur Ermittlung und Überprüfung (Simulation) von Sprachregeln bzw. Strategien der Analyse und Übersetzung. Auch wenn dabei (derzeit) weniger ein "psychologisch fundierter" Ansatz verfolgt wird (hierzu kann jedoch auf Konzepte des Teilprojekts A3 verwiesen werden /5/), sondern Fragen der prinzipiellen "Machbarkeit" bzw. Alternativverfahren zum menschlichen Vorgehen im Vordergrund stehen, bildet das allgemeine wissenschaftliche Interesse die Grundlage der Forschungen.

Modellentwicklung:

Da - nach der Saarbrücker Konzeption - aus projektökonomischen wie auch technischen Gründen ein geeignetes Werkzeug nötig erschien, wurde schrittweise die sog. SUSY-Basissoftware entwickelt, in erster Linie also als Service-Funktion (ähnlich der Programmiersprache COMSKEE). Je weiter sich die Strategien und Verfahren (z.T. durch Erprobung verschiedener Wege) als machbar erwiesen, um so mehr wurden die Softwarepakete zu einem Bündel geschnürt, um so konsistenter entwickelte sich ein Modellsystem zur maschinellen Sprachanalyse und Übersetzung. Die Software behielt ihre Rolle als Entwicklungsinstrument

in der Grundlagenforschung, zugleich wurde jedoch das Modell an sich zu einem Faktor, der zu Überlegungen Anlass gab, konkretere Ziele bzw. Anwendungen anzugehen.

Ein erstes Anwendungsfeld sollte die maschinelle Indexierung sein. Hierzu musste die SUSY-Basissoftware um Softwarepakete bzw. -teile ergänzt werden; z.T. waren dabei grundlegende Fragestellungen (z.B. nach der Verwertung von Analyseergebnissen, nach der Qualität und Brauchbarkeit der Verfahren) zu stellen. Eine derartige Modellentwicklunq wurde in Regensburg im Rahmen des vom BMFT geförderten Projekts JUDO an exemplarischen Daten (Texten zum Bereich Datenschutz) durchgeführt.

Laborversuche:

Auf der Grundlage der Modelle zur maschinellen Übersetzung und Texterschließung werden inzwischen eine Reihe von Laboranwendungen durchgeführt. Ein wesentliches Unterscheidungsmerkmal zu den Modellen ist - nach dieser Strukturierung - die unmittelbare Einbeziehung von (realistischen) Anwendungssituationen. Während die allgemeine Modellentwicklung noch von konkreten Anwendungssituationen notwendig abstrahiert (auch wenn - z.B. bei JUDO - regelmäßige und intensive Anwenderkontakte vorlagen) werden in den Laboranwendungen bereits Rahmensetzungen von Anwendersituationen eingebracht, auch um den Testanwendern eine "praxisnahe" Einschätzung der prinzipiellen Verwendbarkeit zu geben. Die Systementwicklungen SUSY-BSA (vorausgesetzt wird die Übersetzerumgebung des Bundessprachenamts), SUSY-DJT (vorausgesetzt wird ein Benutzer einer fremdsprachigen Datenbank ohne Kenntnisse der Ausgangssprache) und CTX (Aufbau von Test-Datenbanken mit maschinell erschlossenen Freitextdeskriptoren) sind hier einzuordnen.

Pilotanwendungen:

Bei Pilotanwendungen (bzw. prototypischen Anwendungen) werden Anwender bereits unmittelbar "beteiligt". Angesichts der Komplexität der Probleme einer maschinellen Sprachdatenverarbeitung ist dies eine besonders kritische Phase. Während die vorausgehenden Phasen (des Transfers) noch allein von den Forschergruppen bestimmt waren (sieht man einmal davon ab, dass Forschung Geld kostet und jemand - hier der BMFT - dazu motiviert sein musste, einen derartigen Transferprozess einzuleiten) hängt der weitere Fortschritt (bzw. Umsetzungsprozess) von der Entscheidung des (bzw. der) Anwender(s) wesentlich ab. Diese Problematik war Grund genug, den Prozess als solchen abzusichern, d.h. durch wissenschaftliche Begleituntersuchungen (Projekt TRANSIT, gefördert durch den BMFT) zu flankieren, um nicht aufgrund unzureichender Berücksichtigung der Probleme bei Technologie-Transferprozessen das Projektziel zu gefährden.

Die jetzigen Anwendungen des CTX-Systems befinden sich in einer ersten Stufe der Pilotanwendung (Kooperation mit dem Deutschen Patentamt, dem Fachinformationszentrum Werkstoffe und dem Wissenschaftszentrum Berlin). Um eine erste Stufe handelt es sich deshalb, weil noch kein Übergang in die Selbstanwendung beim Anwender erfolgt ist. Angesichts der Vielzahl technischer, rechtlicher und finanzieller Probleme ist dies ein Schritt, der nur behutsam angegangen werden kann.

Versucht man den gesamten F&E-Prozess zur maschinellen Sprachdatenverarbeitung an der Universität des Saarlandes einmal graphisch zusammenzufassen, so ergibt sich etwa folgendes Bild:

Stufe	Thema		
Grundlagen-forschung	SFB 100: Empirische Sprachforschung mit Computer		
Modell-Entwicklung	Modell SUSY		
	Submodell "Übersetzer-unterstützung"	Submodell "automat. Informativ Übersetzung"	Erweiterung "Indexierung"
Laborsystem	SUSY-BSA	SUSY-DJT	CTX
Pilotanwendung	...	...	Patente
Anwendungen	...	...	...

2.2 Ansätze eines neuen Transfer-Modells

Die vorgestellten "Stufen" des Transfers stellen allenfalls grobe Gliederungen dar, die in der Praxis (auch der vorgestellten Projekte) häufig genug durchbrochen werden. Dies ist notwendig, wenn z.B. Fragen auftreten, die nur durch einen "Rückgriff" auf Verfahrensschritte früherer Stufen (Feedback) geklärt werden können. Andererseits können auch dadurch Probleme auftreten, die bewirken, dass die Merkmale eines Projekts sich im Verlaufe der eigentlichen Arbeit aufgrund neuer Erkenntnisse deutlich verschieben. (Dies ist z.B. bei SUSY-BSA der Fall, wo inzwischen ein großer Nachholbedarf bezüglich einer Modellentwicklung eines Übersetzer-Arbeitsplatzes festgestellt wurde.)

Insbesondere aufgrund der großen Komplexität der Probleme (neue Computer-Technologien, Entwicklung der Telematik, Bedarfsorientierung der potentiellen Anwender) kann man daher diese "äußerlichen" Stufen des Transfers nur als ein Gerüst betrachten. Der Feedback-Komponente des Transfer-Prozesses sollte bei einem Transfer-Konzept weitaus mehr Bedeutung zukommen als dies heute der Fall ist. Ihr könnte u.E. am ehesten durch ein Verfahren Rechnung getragen werden, bei dem entwicklungsrelevante Informationen nicht wiederum über die verschiedenen "Stufen" des Transfers gleichsam "zurückgereicht" (also vom Anwender an den Piloten, von diesem an das Laborsystem usf.) sondern in eine "Expertenrunde" eingebracht werden, bei der gleichsam alle Entwicklungsstufen vertreten sind. Dies könnte ein neues - vielleicht alternatives - Transfermodell ergeben: mit eigenen Fragestellungen, z.B. dem Problem der Entwicklung einer Kommunikationsform, die ein thematisches "Verstehen" bei allen Beteiligten sicherstellt.

Angesichts der gewachsenen Konzentration von Forschung und Entwicklung in dem Bereich der automatischen Verarbeitung natürlicher Sprache an der Universität des Saarlandes bietet sich eine reelle Chance, ein derartiges Modell zu erproben, das alle "Betroffenen" des Transferprozesses (als eines Prozesses der Wechselwirkung) - den (Grundlagen-)Wissenschaftler, den "Ingenieur"/Vermittler wie den Anwender/Kunden - an diese Expertentafel bringt. Hierzu gilt es allerdings, diese auf einer Meta-Ebene angesiedelte informationswissenschaftliche Vorstellung einmal praktisch umsetzen.

Die vorhandene Infrastruktur an der Universität des Saarlandes bietet dafür eine Chance. Dies setzt jedoch voraus, dass nach Möglichkeit die bestehenden Kapazitäten (z.T. reichen die Anfänge der Forschung schon über 20 Jahre zurück) rechtzeitig durch eine angemessene Institutionalisierung gesichert werden. Der Wille aller Betroffenen an der Universität des Saarlandes, sich in ein derartiges Gesamtkonzept einzubinden, kann vorausgesetzt werden. Angesichts der wachsenden Bedeutung der maschinellen Bearbeitung natürlicher Sprache für den internationalen Wissensaustausch wäre eine derartige Investition ohne Zweifel von großem nationalem wie internationalem Wert.

Anmerkungen und Literatur

/1/ Vgl. H. Eggers et al: Elektronische Syntaxanalyse der deutschen Gegenwartssprache, Tübingen 1969.

/2/ Vgl. den Projektbericht: H. Zimmermann, E. Kroupa, G. Keil: CTX - Ein Verfahren zur Computergestützten Texterschließung (BMFT-FB-ID 83-006), Karlsruhe 1983.

/3/ Dieses Projekt ist entstanden im Rahmen eines Kooperationsabkommens zwischen der Bundesrepublik und Japan. Ziel ist es, die (natürliche) englische Sprache als "Switching Language" zu nutzen. Dabei übersetzt ein in Japan entwickeltes Übersetzungssystem vom Japanischen ins Englische und ein in Deutschland entwickeltes System (SUSY) vom Englischen ins Deutsche (und umgekehrt).

/4/ Vgl. hierzu z.B. die allgemeine Einführung in G. Habicht/ H. Kück: Bedeutung und Arbeitsweisen von Technologie-Transfer-Einrichtungen in der Bundesrepublik Deutschland, Göttingen 1981.