23.09.2005, 11.00 Uhr, Geb. C 7.4, R.1.17

15. EDV-GT: Arbeitskreis Maschinelle Übersetzung

Referenten: Herr Vorsitzender Richter am Bundespatentgericht Dr. *Wolfgang* Tauchert
Herr Universitätsprofessor Dr. *Manfred* Pinkal

Der Arbeitskreis wurde von Herrn Dr. Tauchert geleitet, der zur Zeit vorsitzende Richter am Bundespatentgericht ist. Herr Professor Manfred Pinkal referierte zum Thema Maschinelle Übersetzung und schöpfte dabei unter anderem aus seinem Erfahrungsschatz aus dem Verbmobil-Projekt, das 2001 den Preis des Bundespräsidenten für Technik und Innovation erhielt.

Einführend stellte Herr Tauchert vom Bundespatentgericht dar, dass es sich bei der maschinellen Übersetzung um ein Spezialgebiet der Sprachtechnologie für Computerlinguistik handelt, worin die Universität des Saarlandes zu den fünf weltbesten Standorten neben Edinborough und anderen zählt. Allerdings wird die Maschinelle Übersetzung zurzeit, nach 8 Jahren aktiver Recherche, in Saarbrücken nicht mehr betrieben.

Professor Pinkal stellte danach die fünf Fragen, an denen sich der folgende Vortrag orientieren sollte:

- Können Computer überhaupt übersetzen?
- Was sind die Probleme die bei maschineller Übersetzung auftreten können?
- Welche Lösungen sind möglich?
- Was unterscheidet die maschinelle Übersetzung von der *maschinengestützten* Übersetzung (Mensch-Computer-Interaktion)?
- Was gibt es sonst in diesem Bereich?

1. Können Computer übersetzen?

Diese Frage wurde mit einer kleinen Anekdote eingeleitet: Mit dem Sputnikschock hat man vor allem in den USA bemerkt, dass Russland in gewissen Bereichen sehr gut war. So wollte man sich deren Erkenntnisse zunutze machen und setzte zum ersten Mal Computer ein, um eine Übersetzung vom Russischen ins Englische zu unterstützen und zu erleichtern. Allerdins sprach der Abschlussreport ein vernichtendes Urteil, was dazu führte, dass die maschinelle Übersetzung in den USA im Keim erstickt wurde und die nächsten 15 Jahre überhaupt nicht weiter verfolgt wurde.

Heute wird dieser Bereich immer weiter erforscht um die Prozesse zu optimieren. Es gibt sogar frei zugängliche Übersetzungsangebote im Internet, wie beispielsweise Babelfish(http://babelfish.altavista.com), den Altavista Standardübersetzer, der seinen Namen dem Buch « Per Anhalter durch die Galaxie » von Douglas Adams entlehnt hat. Allerdings kommt man mit diesem Programm auch schon bald an die Grenzen des Möglichen. Gibt man zum Beispiel den Text « Über allen Gipfeln ist Ruh ...» (http://www.goethe-net.de/nachtl.htm) aus Goethes Gedicht « Wanderers Nachtlied » ein und lässt ihn ins Englische übersetzen, so kann man mit einiger Fantasie noch Sinn aus dem Ergebnis schöpfen. Gibt man nun diesen englischen Text wieder ein und lässt ihn auf Deutsch übersetzen, sollte man erwarten, wieder Goethes Gedicht in deutscher Sprache vor Augen zu haben. Doch hier stellen sich die Probleme ein. Denn das Ergebnis ist in vielen Teilen weit vom Deutschen Original entfernt. An einigen Beispielen wurde dies erläutert:

Deutsch (Goethe)	***Englisch(Babelfish)***	***Deutsch (Babelfish)***
Ruh	rest	Rest

Der Grund für diese Divergenz liegt in der Tatsache, dass es für die Worte in Quell- und

Zielsprache viel zu viele mögliche Verbindungen gibt. Dies wird deutlich, wenn man bei dem Online-Wörterbuch Leo (http://dict.leo.org/) « Ruhe » eingibt und ins Englische übersetzen lässt. Das Ergebnis sind 20 verschiedene Alternativen. Lässt man « rest » wieder ins Deutsche übersetzen, so erhält man wiederum 6 verschiedene Alternativen. Hier wird klar, warum das maschinelle Übersetzungsprogramm in der deutschen Babelfish-Version nicht wieder zum ursprünglichen Wortlaut des Goethe-Gedichts zurückfinden konnte. Einige weitere Beispiele für diese Übersetzungsdivergenzen in folgender Tabelle:

Deutsch (Goethe)	***Englisch (Babelfish)***	***Deutsch (Babelfish)***
Du	you	Sie
Hauch	breath	Atem
Keinen	do not	Nicht (kein Übersetzungs-äquivalent im Deutschen)

All diese Beispiele sind auf Probleme des « mismatch » (zu Deutsch « Fehlanpassung »), der Mehrdeutigkeit, der Einwort-Mehrwortausdrücke oder der unterschiedlichen Reihenfolge zurückzuführen.

Nach diesen Erläuterungen wird klar, dass eine vollautomatische, hochwertige Übersetzung für Dichtung und rechtlich relevante Dokumente in absehbarer Zeit nicht im Bereich es Möglichen liegen wird. Allerdings ist die maschinelle Übersetzung durchaus für eine Reihe von Zwecken anwendbar.

2. Probleme im Detail

Um die Probleme im Einzelnen darzustellen, wurde im nächsten Programmpunkt das Verbmobil-Projekt vorgestellt, das vom Bundespräsidenten mit einem Preis Preis für Technik und Innovation ausgezeichnet wurde (http://www.dfki.de/zukunftspreis). Dieses Projekt im Bereich der Sprachtechnologie ist weltweit das größte seiner Art. Über acht Jahre haben 120 Wissenschaftler an vielen Standorten im Zeitraum von 1992 bis 2002 daran gearbeitet. Dabei sollte die Dialogübersetzung automatisiert werden. In einer face-to-face Situation wurden mit Mikrofon, beziehungsweise Telefon Gespräche über Termin- und Reiseplanung aufgezeichnet und verwertet. Übersetzungssprachen waren Deutsch, Englisch und Japanisch, wobei der Sprachumfang bei ersteren 10.000 Worte und bei letzterer 2.500 Worte maß. Anhand des Verbmobil-Projekts wurden die Probleme, die bei maschineller Übersetzung auftreten können, im Folgenden erläutert.

Lexikalische Mehrdeutigkeit
In der Quellsprache muss das Wort gefunden werden, das tatsächlich gemeint ist.

Deutsch (Verbmobil)	***Mögliche Übersetzungen im Englischen***
Termin	– appointment (= Terminzeitpunkt) – time slot (= Zeitintervall)
gehen	– move – act – feel – last

Da auf diese deutschen Worte mehrere Englische Übersetzungen zutreffen können, kann man sie al

lexikalisch mehrdeutig bezeichnen. Um aus diesen möglichen Lesarten die richtige herauszufinden, muss der Kontext analysiert werden. Suboptimal wäre es, dabei nur auf die fünf Worte vor und hinter dem zu übersetzenden Wort zu achten. Idealiter sollte dabei die Syntax mit Subjekt, Verb und Objekt herangezogen werden. Weiter sollte auch der satzinterne Kontext Beachtung finden. Beispielsweise können die Worte « before » und « in front of » auf eine zeitliche beziehungsweise örtliche Verwendung des deutschen Wortes « vor » hinweisen. Problematisch wird es auch dann wieder, wenn die Umgangssprache für Ungenauigkeiten sorgt, wie bei dem Satz « Wir treffen uns *vor* Hamburg », was hochsprachlich ausgedrückt heißen sollte « Wir treffen uns vor dem Meeting/ der Messe/ der Sitzung etc. *in* Hamburg. ». Das « vor » hat in dem ersten Satz also zeitliche Bedeutung. Da es jedoch im Satz vor der Ortsangabe « Hamburg » steht, könnte der Computer es als örtlich einstufen und somit falsch übersetzen.

Referentielle Mehrdeutigkeit

Probleme der referentiellen Mehrdeutigkeit sind inhärent mehrdeutige Pronomen, die nicht eindeutig maskulin, feminin oder plural sind. In diesem Fall sind andere Informationen nötig.

Bsp.: « **Peter** liebt seinen Hund, obwohl **er** ihn manchmal beißt. »
Hier sind Parallelitäten im Satzbau festzustellen, die zu falscher Interpretation führen könnten. Davor sind Filtermechanismen nötig, die Bedeutungs-, Kontext- und Weltwissen mühelos und in Echtzeit analysieren können.

Bsp.: « In der **Zunkunft** werden **Maschinen** erfunden werden, die immer mehr auf ihre **Umwelt** reagieren und in der Lage sind **ihren** Betrieb an wechselnde Bedingungen anzupassen. »
Dabei kann sich « ihren » sowohl auf « Zukunft », auf « Maschinen » als auch auf « Umwelt » beziehen. Hier kann der globale Kontext Aufschluss geben, der in Form von Ablaufschemata, so genannten scripts, mit berücksichtigt werden muss. So ist es auch bei dem Satz « Geht es bei Ihnen? », wobei « bei Ihnen », je nach Kontext mit « at your place » oder mit « for you » übersetzt werden kann.

Sprachen haben unterschiedliche Granularität

Aufgrund dieser untschiedlichen Granularität von Sprachen gibt es für ein Wort in der Quellsprache oft mehrere mögliche Worte in der Zielsprache. Hier einige Beispiele für Granularitätsunterschiede sowohl bei der Übersetzung vom Englischen ins Deutsche, als auch vom Deutschen ins Englische.

Englisch (Verbmobil)	***Deutsch (Verbmobil)***
I am going to Hamburg. ==►	– fahre – fliege
– change train – change plain	◄== umsteigen
doctor (geschlechtsneutrale Berufsbezeichnung)==►	– Arzt – Ärztin
– I am going – I will go	◄== Ich fahre nach Hamburg.

Will man nun deutsche oder englische Sätze ins Japanische übertragen, wird man sich mit der Problematik auseinandersetzen müssen, dass in Japan die Höflichkeitsformen durch Verbpräfixe oder -suffixe ausgedrückt werden. Daher ist in diesem Bereich besonders viel intellektuelle Arbeit bei der Übersetzung nötig.
Ein anderes Problem ergibt sich daraus, dass das Japanische keine Artikel kennt. Demnach muss man sich bei der Übersetzung ins Deutsche oder ins Englische entscheiden, ob man nun bestimmte oder unbestimmte Artikel einfügt.

Des weiteren werden im Japanischen häufig ganze Satzteile – wie Subjekte und Objekte - tendenziell weggelassen, wenn sie aus dem Kontext erschließbar sind. Dieses Phänomen der « Null-Anapher » wurde an folgendem Beispiel demonstriert. Im Japanischen heißt es einfach « Termin ausgemacht? ». Im Deutschen wäre hier fragwürdig ob « er mit Ihnen » oder « Sie mit ihm ». Diese Unterscheidung kann man im Japanischen nur aus den unterschiedlichen Höflichkeitsformen schließen, die bei direkter Anrede gebraucht werden.

An einem Beispielfilm des DFKI (http://www.dfki.de/web/) wurde weiterhin erläutert, dass auch die akustische Satzmelodie ausgewertet werden muss. Außerdem ist bei Sprachanalyse auch ein Worthypothesengraph hilfreich, der Selbstkorrekturen des Sprechers beim Sprechen richtig berücksichtigt.

Um der unterschiedlichen Granularität der Sprachen Rechnung zu tragen hatte schon der französische Wissenschaftler Vauquois das Vauquois-Dreieck beschrieben. Auch us-amerikanische Wissenschaftler haben 1957 einen ähnlichen Versuch unternommen, scheiterten aber damit. Das beste Modell in diesem Bereich ist wohl das Inerlingua Modell. Denn bei der Analyse wird eine Ebene der Bedeutungsrepräsentation zwischen die eine und die andere Sprache geschaltet. Um dieses Verfahren jedoch korrekt durchführen zu können, sind möglichst flächendeckende Grammatiken nötig. Diese sind bislang im Englischen existent, jedoch noch nicht im Deutschen. Problem im Deutschen sind hierbei auch die trennbaren Präfixverben, wie « auffordern », das zu « fordern ... auf » werden kann. Daher wurde eine Strukturierte Ein- und Ausgabe generiert (Analyse – Transfer – Generierung).Înterlingua bedient sich einer sprachunabhängige Interlingua-Darstellung. Diese sieht auf den ersten Blick sehr gut aus, funktioniert nur leider nicht, weil es viel zu umständlich ist, alle Sprachen einzubeziehen. Beim Transfer-Modell hingegen wird mit Transfermodulen gearbeitet. Hier ist jedoch der Nachteil, dass man mit jeder weiteren Sprache immer mehr Module benötigt.

4. Computer übersetzen

Im Fazit wurde festgestellt, dass Computer heute keine hochwertige Übersetzungsleistung liefern können. Jedoch ist es möglich und sinnvoll Gebrauchstexte mit maschineller Übersetzung approximativ übersetzen zu lassen. Dies wurde anhand der Internetseite der japanischen Industrial Property Digital Library des japanischen Patent Office (www.ipdl.ncipi.go.jp/homepg_e.ipdl)gezeigt.

Webdienste in diesem Bereich sind daher sinnvoll und hilfreich, um beispielsweise interessante von uninteressanten Texten zu unterscheiden erkennen zu können, ob eine richtige, fachmännische Übersetzung eventuell nötig oder sinnvoll ist.

Alternativen zur maschinellen Übersetzung

Ein weiteres Verfahren ist wissens- und regelbasiert, das sich Lexika für Quellsprache und Zielsprache zunutze macht. Darüber hinaus gibt es auch statistische Verfahren – angewendet von Philipps Aachen. Basis ist eine Menge von Textkorpora aus den Akten des europäischen Parlaments, die sehr interessante Anhaltspunkte für das Alinieren – die Satzzuordnung – bieten, da alle Urteile sich im Satzbau gleichen. Außerdem gibt es auch die Übersetzung am Beispiel, genannt « Translation by example ». In Form von webbasierter Übersetzung funktioniert diese sehr gut.

Das Institute for information science in Amerika benutzt vor allem für die Übersetzung vom Arabischen ins Englsiche Anwendung von wissens- , regel- und webbasierten Verfahren, die eventuell bald auch Babelfish übertreffen.

Aber auch der Mensch kann beteiligt werden um alles zu optimieren. Ein Beispiel dafür sind die HAMT – Systeme. Unter ihnen ist « Systran » - in Babelfish eingebunden - das Beste. Andere

HAMT – Systeme sind der « Linguatec Personal Translator », den IBM verwendet (http://idw-online.de/pages/de/news127548), oder auch die Opensource-Software « Logos », die unter www.trados.com jedermann zugänglich ist.

Was gibt es sonst noch?

Das Projekt CORTE (http://www.coli.uni-saarland.de/projects/corte/) bietet beispielsweise Informationszugriff und Terminologieerweiterung. Definitionen werden hierbei mit sprachtechnologischen Mitteln erstellt - Definiens, Definiendum, notwendig oder hinreichend. Dies geschieht in Kooperation mit dem Institut für Rechtsinformatik von Professor Herberger (http://rechtsinformatik.jura.uni-sb.de/) mit dem Zugriff auf die juris-Datenbank (http://www.jurpc.de/index.html).
Des weiteren wird an der Dialogsteurung im Fahrzeug gearbeitet - ein Projekt namens NaDia-System mit BMW. Dies ermöglicht Navigation, Handy und Voice-Browsing für Newsticker in einem und wird eventuell in der neuen Siebener-BMW-Serie in drei bis vier Jahren eingebaut.

Fragen

Ein Zuhörer wies auf die Problematik hin, dass deutsche jurisitsche Texte, die auch mit Algorythmen dargestellt werden können, aufgrund der unbegreiflichen Ausdrucksweise von Richtern selbst in der Quellsprache unverständlich werden. Es kam dabei die Frage auf, ob man in der maschinellen Übersetzung auch von Sprachlexika profitieren könne. Dies sei möglich, je stärker man sich im jeweilgen Fachbereich bewegt, desto wichtiger werden die Lexika. Damit wird dann auch das Problem der Mehrdeutigkeit besser bewältigt.
Weiterhin wurde die Frage aufgeworfen, wie vorgegangen wird, wenn ein Wort in der Zielsprache nicht existiert. Beispielsweise gibt es im Japanischen kein Verb für « Trinken ». Allerdings existiert ein Verb, das aber auch manchmal trinken und einnehmen von Medikamenten in flüssiger Form umfasst. In diesem Fall wird dann dieses verwendet.
Es kam auch die Frage auf, wie gut Übersetzer in Kugelschreiberform sind. Diese wurden vom Referenten als uninteressant eingestuft, da sie kein sehr großes Gedächtnis haben können und daher nur auf Wortebene übersetzen. Vermutlich sind auch nur kleine Wortgruppen übersetzbar. Das Problem dabei liegt im begrenzten Speicher, der auch beim BMW-Projekt Grenzen setzte. So wurden dabei nur 1.500 bis 2.000 Worte und Orts- und Sraßennamen, insgesamt 50.000 Begriffe verwendet. In diesem Zusammenhang wurde auf das deutsche Wahrich-Wörterbuch von Bertelsmann (Wahrich-Text-Corpus-Digital) hingewiesen, das 1 Mrd. Textwörter enthält. Auch wurde angeregt, dass die Internetseite des Europäisches Patentamt (http://www.european-patent-office.org) auch interessant sei. Außerdem wurde auf « Thesaurus » (http://thesaurus.com/) hingewiesen, das die Zusammenführung juristischer Begriffe unternimmt, die gleiches bedeuten. Dabei wurden auch die Kontakte zu iuris (http://www.juris.de) erwähnt.

Elisabeth Drechsel