Zeit Wissen Germany March 2006

VERLOREN IN ÜBERSETZUNG

▶Wenn der Computer übersetzt, hat der Mensch was zu lachen – die Programme liefern manchmal verständliche, selten korrekte Texte. Google-Programmierer tüfteln nun an einem neuen Konzept. Sie füttern die Festplatte mit Unmengen übersetzter Texte und lassen den Computer tun, was er nun einmal am besten kann: rechnen.

TEXT KARSTEN LEMM

Kennzeichnen Sie Twain, der berühmte Schriftsteller, klagte einmal: ›In Paris starrten sie einfach an, als ich mit ihnen auf französisch sprach; Ich nie folgte, mit, jene Idioten zu bilden, verstehe ihre Sprache.‹«

Na, das übersetzen wir besser noch mal; die Übersetzungs-Software der Firma Systran ist da augenscheinlich überfordert. Sie kennt nicht mal Mark Twain und versucht prompt, den Vornamen mit zu übersetzen – so wird in der Übersetzung vom Englischen ins Deutsche aus *Mark* dann »Kennzeichnen Sie«. Tatsächlich hat Twain dies gesagt: »In Paris haben sie mich immer nur angestarrt, wenn ich Französisch mit ihnen gesprochen habe; nie ist es mir gelungen, den Idioten ihre eigene Sprache verständlich zu machen.«

Systran gehört zu den führenden Anbietern in Sachen Maschinenübersetzung, die Software wird von der Europäischen Union ebenso eingesetzt wie von Daimler, Cisco und anderen Großunternehmen. Und doch

ist das System schon mit einem simplen Ausdruck wie *Well done!* überfordert. Daraus macht es fröhlich »Brunnen getan!«, schließlich kann *well* im Englischen beides bedeuten, »gut« oder »Brunnen«, je nach Zusammenhang, und den hat die Software offenbar nicht erkannt.

Es wäre unfair, die mangelnde Sprachkompetenz der Computer auf inkompetente Programmierer zu schieben. Die Übersetzungspannen offenbaren nur, was Sprachforscher schon lange wissen: Sprache ist eine hoch komplexe Angelegenheit, besonders für ein schlichtes Rechenhirn, dessen Welt bei null beginnt und bei eins aufhört.

Seit über 30 Jahren bemüht sich Systran, Computern die Feinheiten menschlicher Kommunikation beizubringen. In mühevoller Kleinarbeit füttern Programmierer in Paris Tag für Tag ihre Rechner mit Regeln für derzeit 36 Sprachpaare, von Englisch/Chinesisch bis Portugiesisch/Französisch. Grammatik, Satzbau und Bedeutung werden dabei berücksichtigt, ergänzt um einen lexikalischen Wortschatz. Die Idee: Beschreibt man dem Computer die Sprachen nur ausführlich genug, müsste er Texte übersetzen können, ohne ihren Sinn zu kennen. Erfüllt hat sich die Hoffnung bisher nicht.

»Übersetzen ist ein schwieriges Problem«, sagt Systran-Chef Dimitris Sabatakakis. »Computer besitzen nun mal keinerlei Intelligenz. Ein Mensch kann eine Sprache schon mit tausend Wörtern sprechen. Wir haben Computermodelle mit über einer Million Wörtern – und trotzdem können sie nicht mit einem Menschen mithalten.« Aber besser eine radebrechende Übersetzung als gar keine, argumentiert er. »Es gibt zu viele Texte, die übersetzt werden müssen. Menschen allein können das gar nicht schaffen.«

Besonders im Internet wächst der globale Informationshunger, und fast alle großen Portale von AOL bis Yahoo! bieten Übersetzungshilfen an, die auf Systran-Software basieren. 30 Millionen Web-Seiten übersetze seine Firma Tag für Tag, sagt Sabatakakis, also müssen es die Menschen doch nützlich finden, oder? »Wir zwingen ja niemanden, unser System zu nutzen.«

EIGENTLICH SOLLTE die Sache längst erledigt sein. Als IBM am 7. Januar 1954 zum ersten Mal einen Automaten vorführte, der ein paar Sätze aus dem Russischen ins Englische übertragen konnte, gab sich der Projektleiter siegessicher: Schon in »fünf, vielleicht drei Jahren« dürfe man von Rechenmaschinen brauchbare Dolmetscherdienste erwarten, verkündete Georgetown-Professor Leon Dostert den versammelten Reportern. Eilfertig jubelte die *New York Times* tags darauf, das neue System werde Übersetzungen »schnell, sorgfältig und einfach« machen.

Als es dann weder schnell noch einfach ging, kam der Spott. Hartnäckig hält sich die Legende, einmal habe ein Computer den Satz »Der Geist ist willig, aber das Fleisch ist schwach« aus dem Englischen ins Russische übertragen sollen – wobei dann herauskam: »Der Wodka ist gut, aber das Steak ist lausig.« Da lacht der Forscher. Schön wär's, wenn automatische Übersetzer tatsächlich so zielsicher daneben lägen. Viel öfter erinnern die Resultate an übersetzte Bedienungsanleitungen für koreanische Stereoanlagen oder Fotoapparate aus den frühen achtziger Jahren – unterhaltsam, aber wenig hilfreich.

»Zurzeit ist Maschinenübersetzung noch unbefriedigend«, räumt Franz Och ein, einer der führenden deutschen Computerlinguisten. »Ich selbst nutze sie nicht.« Jedenfalls nicht die herkömmlichen Dienste, die allgemein im Internet zugänglich sind. Och, 34, ein gebürtiger Erlanger, den es in die Google-Zentrale nach Kalifornien verschlagen hat, bastelt sich lieber seine eigene Übersetzungsmaschine. Sie ist noch längst nicht fertig, und deshalb setzt Google bei seinen »Sprachtools« vorerst noch auf Systran. Aber beim Wettbewerb der automatischen Dolmetscher, den die US-Behörde NIST jedes Jahr veranstaltet, zeigte Ochs System bereits, was es kann: Es ließ alle Konkurrenten, inklusive Systran und diverser Universitäten, mit weitem Abstand hinter sich, gleich beim ersten Anlauf.

Der Unterschied lag in einer gewissen Disziplinlosigkeit, einer radikalen Abkehr von der Strategie, die Computerlinguisten seit Jahrzehnten verfolgen. Während andere weiterhin versuchen, der Rechenmaschine Sprache beizubringen, sagt Och: »Regeln, das ignorieren wir alles!« Denn Regeln machen nichts als Ärger. »Für jedes Sprachpaar braucht man neue Experten, das ist ein sehr zeitraubender und teurer Ansatz. Da gibt es so viele Feinheiten, das ist das entscheidende Problem.« Wirft man aber alle Regeln über Bord, kann man sich auf das konzentrieren, was Computer ohnehin am besten beherrschen: rechnen nämlich.

Schon vor beinahe 20 Jahren kamen IBM-Forscher auf die Idee, ihr kniffliges Sprachproblem mit abstrakter Statistik zu lösen. Nicht auf den Inhalt der Texte sollte man achten, dachten sie sich, sondern auf charakteristische Wortstellungen und immer wiederkehrende Muster.

Vergleicht man große Mengen schon fertiger Übersetzungen, kann man zählen, wie oft bestimmte Wörter neben anderen stehen – etwa »ich« neben »gehe« im Deutschen und entsprechend *je* neben *vais* im Französischen. So analysiert man Satz für Satz, ohne sich um den Sinn zu kümmern. Irgendwann hat der Computer genug Material angehäuft, um frische Texte anzugehen, für die es noch keine Übersetzung gibt. Allein auf Basis der Erfahrungswerte sollte er dann in der Lage sein, für jede Wortkombination und jeden Satz die wahrscheinlichste Übersetzung zu errechnen.

Was den Statistikern lange fehlte, waren die Unmengen übersetzter Texte, ohne die der Lösungsansatz nicht funktioniert. Doch dann kam das Internet. Auf einmal war es möglich, sich dort zu bedienen, wo immer schon ein produktiver Sprachenwirrwar herrschte. »Die Vereinten Nationen sind eine tolle Quelle für Übersetzungen«, sagt Franz Och, »genau wie die EU.« Auch die Web-Seiten internationaler Großkonzerne zapfen die Forscher heute gern an, um ihre Software mit Vergleichsdaten zu füttern –und niemand hat bei der multilingualen Such- und Sammelaktion bessere Karten als Google. »Die Infrastruktur ist fantastisch«, schwärmt Och. »Hier werden acht Milliarden Web-Seiten erfasst und ständig aktualisiert. Die liegen einfach da, und mit denen kann man dann rechnen.«

Der Aufwand ist typisch Google. Für den NIST-Wettbewerb ließ Ochs Team 1000 PCs 40 Stunden lang über der Aufgabe brüten und jeweils 100 Nachrichtentexte aus dem Arabischen und Chinesischen ins Englische zu übersetzen. Dabei hantierten die Rechner mit einem Datensatz von 200 Milliarden

»DA SIE IHR BETT, ALSO BILDEN, SIE MUSS IN IHM LIEGEN.«
(AS YOU MAKE YOUR BED, SO YOU MUST LIE IN IT.)
ÜBERSETZUNG: ALTAVISTA-BABELFISH

Wörtern – »das größte Sprachmodell in der Geschichte der Menschheit«, wie sein Schöpfer Franz Och stolz vermerkt.

Die Mühe wurde belohnt: mit verständlichen Sätzen statt des üblichen Kauderwelschs. »ElBaradei: Inspektoren brauchen ›einige Monate‹, um ihre Aufgabe zu erfüllen«, lautete zum Beispiel die (englische) Übersetzung, die das Google-System für eine arabische Schlagzeile lieferte. Eines der Konkurrenzmodelle, das nach dem traditionellen Regelprinzip vorging, machte aus dem gleichen Satz: »Der Bradi: Die Inspektoren brauchen zu ›einige Monate‹ für Ende wichtige ihr.«

»DA IST EIN gutes System entstanden«, lobt Alex Waibel, Professor für Computerwissenschaft in Karlsruhe und Pittsburgh. Der 49-Jährige träumt schon sein halbes Leben lang davon, einen Universalübersetzer zu entwickeln, bei dem man auf der einen Seite reinspricht, »und auf der anderen Seite soll's in einer anderen Sprache rauskommen«. Das hatte er sich zwar um einiges einfacher vorgestellt, aber jetzt, nach vielen Rückschlägen, wittert er Morgenluft. »Das statistische Modell ist haushoch überlegen«, schwärmt Waibel. »Wenn Sie fast beliebig viele Daten sammeln und auswerten können, wächst die Übersetzungsfähigkeit automatisch, das System lernt immer weiter. Das sehe ich als großen Durchbruch.«

»ALLER QUELLEN HERVOR, DASS ENDEN HERVORQUELLEN.« (ALL'S WELL THAT ENDS WELL.)
ÜBERSETZUNG: ALTAVISTA-BABELFISH.

Waibel ist ein quirliger Mann mit hoher Stirn und Vollbart, der dem technischen Fortschritt gern ein paar Schritte voraus ist. An seinem Institut an der Carnegie Mellon University arbeiten Mitarbeiter an einer Art Schallwellen-Kanone, die Töne ganz gezielt auf einzelne Personen fokussieren soll. »Stellen Sie sich vor, die Vereinten Nationen hätten so etwas«, sagt Waibel, »jeder bekäme seine eigene Übersetzung zugebeamt.« Ein anderer Mitarbeiter hat ein System entwickelt, das die Bewegungen der Gesichtsmuskeln beim Sprechen erfasst und das Gesagte sofort übersetzt. »Die Elektroden wären in Ihre Wangen implantiert«, schwärmt Waibel, »dann könnten Sie in jeder Sprache sprechen, die das System unterstützt.«

Zugegeben, das ist noch Science-Fiction. Aber das war die Erfindung, die Waibel als Nächstes vorführt, vor ein paar Jahren auch noch. Vor einer Diawand in einem kargen Konferenzraum setzt er ein Headset auf und spricht gewichtige Worte in das Mikrofon: *»The mission of our center is to produce technology that makes communication between the people of the world easier«*, sagt Waibel, und während er redet, wirft ein Projektor die deutsche Übersetzung hinter ihm auf die Leinwand. »Die Mission für unser Center«, steht da Sekundenbruchteile später, »ist zu produzieren Technologie das macht die Kommunikation zwischen Menschen der Welt zu erleichtern.«

Perfekt ist das nicht, aber im Vergleich zu früheren Versuchen ein großer Schritt voran. »Das Problem mit spontaner Sprachübersetzung war immer: Man hat ein uneingeschränktes Vokabular und oft auch ein ziemliches Gestotter«, erklärt Waibel. Schon das Verstehen der Ausgangssprache ist eine Herausforderung, »ein Übersetzungsschritt zwischendrin«, weil kaum jemand druckreif spricht. Oft fehlen Wörter oder halbe Sätze, der Sprecher bricht ab und setzt neu an. Außerdem wird genuschelt und gemurmelt, bis der Forscher sich die verbleibenden Haare rauft. »Am schlimmsten sind Gespräche und Konferenzsituationen«, sagt Waibel, leidgeprüft durch jahrelange Mitarbeit am deutschen »Verbmobil«-Projekt, bei dem es um das Erfassen und Übersetzen gesprochener Sprache ging. Deshalb konzentriert sich seine Forschungsgruppe bei ihrem System vorerst auf einzelne Menschen, die Vorträge halten.

Schon das ist schwer genug, wie sich zeigt, als Waibel anfängt zu improvisieren. *»Can you give me a hand please«*, sagt er – können Sie mir bitte helfen. Der Computer macht daraus: »Können Sie mir eine Seite bitten.« Ein kurioser Fehler, erklärlich vielleicht dadurch, dass das Dolmetscherprogramm mit Debatten aus dem EU-Parlament trainiert wurde. »Es kann sein, dass der Ausdruck *on the other hand* dort besonders häufig vorkommt«, spekuliert Waibel. Zusammenhängende Begriffe bereiten den Forschern ohnehin Kopfzerbrechen, weil ihre Software sich beim Analysieren der Ausgangsdaten derzeit nicht den ganzen Satz anschaut, sondern immer nur zwei benachbarte Wörter. »Damit können Abhängigkeiten wie im Deutschen nicht erfasst werden«, sagt Stephan Vogel, einer von Waibels Mitarbeitern. »Für viele Sprachen ist die Wortstellung das Problem. Wir kriegen viele Wörter richtig hin, aber nicht unbedingt in der richtigen Reihenfolge.«

ABER DAS SIND Kinderkrankheiten. Die meisten statistischen Modelle sind erst wenige Jahre alt, »und in machen Fällen hat man die regelbasierten Ansätze schon überrundet«, sagt Waibel. Sollte die automatisierte Übersetzung also nach 50 Jahren stop and go nun spürbar vorankommen? Die Chancen stehen gut. Das sieht sogar der Mann so, dessen Unternehmen sich bisher ganz auf den traditionellen Lösungsweg konzentriert hat. »Die statistische Methode ist zwar nicht bahnbrechend neu«, sagt Systran-Chef Dimitris Sabatakakis. Neu sei nur, dass man jetzt viele, viele Daten sammeln könne und die Rechenhirne schnell genug seien, die Zahlenmengen auch zu bewältigen. Dennoch: »Das hilft enorm, die Systeme zu verbessern. Wir sehen echten Fortschritt.«

Um nicht zurückzubleiben, machen die Systran-Linguisten sich nun daran, Statistik in ihre Regelwerke einzubauen. Waibel und seine Kollegen planen im Gegenzug, ihre Modelle mit ein paar Grundregeln der Sprache anzureichern, um die Trefferquote ihrer Software zu erhöhen.

Wie gut können die Maschinen werden, die am Ende dabei herauskommen? »Ich glaube nicht, dass Übersetzer um ihre Zukunft bangen müssen«, sagt Google-Forscher Franz Och. »Maschinelle Übersetzung wird nie perfekt werden, nie so gut wie ein menschlicher Übersetzer.« Romane, Gedichte, Liedertexte, alles Literarische und Poetische, »das bleibt dem Menschen vorbehalten«, sagt Dimitris Sabatakakis. Und das ist ja auch gut so, oder? »Stellen Sie sich vor: Gabriel Gárcia Márquez, von einer Maschine übersetzt! Das wäre doch geradezu abstoßend.« Allenfalls bei relativ simplen Sachtexten kann der Systran-Chef sich vorstellen, dass Computer in einigen Jahren »hundertprozentige Qualität« liefern.

Und selbst dann sollte man auf der Hut sein – vor allem bei medizinischen Texten, egal, in welcher Sprache. »Nehmen Sie sich vor Gesundheitsbüchern in Acht«, riet schon Mark Twain. »Sie könnten an einem Druckfehler sterben.« ■