

Peter Matussek

Lesen und lesen lassen

Mit dem Kürzel ICR (intelligent character recognition) will die Branche der Lesemaschinenhersteller signalisieren, daß sie die Legasthenieprobleme des herkömmlichen OCR (optical character recognition) überwunden hat. Doch wie intelligent ist die neue Technik wirklich? Macht sie das mühselige Korrekturlesen überflüssig? In welchem Verhältnis steht der Aufwand zum Nutzen?

"Nun müssen mir meine Freunde noch eine Vorlesemaschine erfinden." Mit diesem Satz beendete anno 1882 Nietzsche seine kurze, aber heftige Liebesaffäre mit einer der ersten Schreibmaschinen, die seinerzeit in Deutschland zu haben waren. Die „Malling Hansen“, die den extrem kurzsichtigen Philosophen einen Sommer lang aus Freude über die wiedererlangte Lesbarkeit der eigenen Produkte zu hymnischen Gedichten beflügelt hatte, war im feuchten Genueser Klima unbrauchbar geworden. Die Mechanik klemmte, das Farbband klebte, der Kopfschmerz ob der überanstrengten, fast erblindeten Augen kehrte zurück ...

Zeitsprung ins Jahr 1974. Nietzsche ist lange tot, aber sein Wunsch nach einer Vorlesemaschine, von dem er sich eine gesteigerte Fortsetzung seines kurzweiligen Tippglücks erhofft hatte, wird endlich erhört. Raymond Kurzweil heißt bezeichnenderweise der Mann, der es sich nun zur Aufgabe macht, ein solches Gerät zu produzieren. Er gründet die Kurzweil Computer Products (KCP) mit der Zielsetzung, einen gedruckten Input in einen phonetischen Output umzuwandeln. Insbesondere Blinde sollen von der neuen "Reading Machine" profitieren.

Zwar ist die Technik der OCR (optical character recognition) zu dieser Zeit bereits erfunden – sie ist im Grunde so alt wie der Computer selbst. Doch die

herkömmliche OCR war beschränkt auf ganz bestimmte, normierte Schriften. Man mußte den Input in der Regel neu eingeben, bevor ihn das Ausgabegerät in Braille stanzen, in Schallwellen modulieren oder auch nur für spätere Weiterverarbeitung elektronisch speichern konnte – ein Handicap, das den Leseunfähigen oder Leseunwilligen von der Aneignung eines Großteils der alltäglichen Drucksachen ausschloß.

Kurzweil will eben diese Legasthenie der OCR-Technik überwinden. Er setzt auf ICR, eine intelligente Buchstabenerkennung, die mit den unterschiedlichsten Schriftbildern und Seitenformaten, ja selbst Druckfehlern (wie sie in der Praxis nun einmal auftreten) zurechtkommen soll.

Knapp ein Jahrhundert nach Nietzsches Utopie ist es soweit: 1978 kommt die Kurzweil Data Entry Machine (KDEM) zum Einsatz. Das Gerät kostet rund 250 000 Mark, was einer heutigen Kaufkraft von gut einer halben Million entspricht und einer großen Verbreitung entsprechend enge Grenzen setzt. Aber die Branche boomt. Der Gesamtumsatz der Texterkennungstechnik liegt in 1986 bereits bei 200 Millionen Mark, und inzwischen dürfte die Milliardengrenze erreicht sein. Die wachsende Bedeutung der ICR läßt sich daran ermessen, daß hier der Preisverfall bei gleichzeitigem Leistungsanstieg noch rasanter ist als in der Computerindustrie überhaupt. Knapp 10 000 Mark kosten die KDEM-Nachfolgesysteme 1990, bei 64facher Speicherkapazität und 25facher Geschwindigkeit gegenüber der Erstausgabe.

Eine rechnerunabhängige Softwarevariante namens AccuText ist inzwischen auf dem Markt, die in der Macintosh-Version etwa 3400 Mark kostet. Was das von der Firma CCS vertriebene Produkt leistet, wollen wir uns nun genauer ansehen, um daran zugleich exemplarisch die Funktionsweise der ICR zu erläutern.

AccuText arbeitet auf der Grundlage der Expertensystemtechnik, einer Anwendung der Künstlichen-Intelligenz-Forschung. Die einprogrammierte "Experten" rücken einer gescannten Druckvorlage in drei Arbeitsschritten zu Leibe: Im ersten Schritt unterscheiden sie Text- und Grafikelemente sowie die einzelnen Wörter voneinander. Was der menschlichen Bildwahrnehmung mühelos und spontan zufällt, Irrelevantes und Wichtiges – etwa Schmutzflecken

und Markierungen – auseinanderzuhalten, ist der binären Logik eines Computers nur durch höchst komplexe Algorithmen beizubringen. Die zuständige Expertengruppe von AccuText berechnet auf der Basis eines Regelwissens, wie hoch die Wahrscheinlichkeit ist, daß bestimmte Bildpunkte zusammengehören, sei es als Buchstabengruppe, als Grafik oder als Spaltenmarkierung, und separiert sie entsprechend für die weitere Analyse.

Jetzt sind die Erkennungsexperten dran. Sie vergleichen jeden einzelnen Buchstaben der separierten Wörter mit Beschreibungen ihrer geometrischen Eigenschaften. Ein kleines e zum Beispiel erkennen sie am oberen Halbkreis, an der horizontalen Linie in der Mitte sowie an der Lage der konvexen und konkaven Krümmungen. Diese Erkennungsmethode richtet sich also nicht nach bestimmten, vorgegebenen Mustern, sondern nach allgemeinen Merkmalen; deswegen spricht man hier im Gegensatz zum pattern matching von feature recognition, die sich sowohl individuellen Schriftbildern durch entsprechendes Training anzupassen als auch fragmentierte Buchstaben zusammensetzen beziehungsweise anstoßende auseinanderzuhalten vermag. Außerdem sind die Erkennungsexperten intelligent genug, Passagen in schlechter Druckqualität mit solchen zu vergleichen, deren Schriftbild weniger Probleme macht, um durch solche Vergleiche Unsicherheiten aufzulösen.

Die auch dann noch auftretenden Unsicherheiten aufzulösen ist der Job des dritten Expertenteams, das sich klassischer Problemfälle wie zum Beispiel der Unterscheidung von g und q oder r und n annimmt. Eine ganz besondere Knacknuß stellen solche Zeichen dar, die in der feature-recognition-Sprache nur als senkrechter Strich zu beschreiben sind. Handelt es sich dabei um ein kleines l, eine Eins, ein großes I oder den unteren Teil des kleinen i? Um solche Ambiguitäten aufzulösen, benutzt AccuText nun lexikalisches Wissen über die jeweilige Landessprache. Dieses besteht zum einen in Wortlisten, die unmittelbar zur Bestätigung dienen können, zum anderen in Auskünften über die Wahrscheinlichkeit bestimmter Buchstabenkombinationen, die etwa das fragliche zweite c in „Schokclade“ als fragmentiertes o identifizieren, weil „kcl“ nun mal im Deutschen nicht vorkommen kann.

Das Kriterium für die Intelligenz eines Systems ist seine Lernfähigkeit. Bei AccuText zeigt sie sich durch recht angenehme Umgangsformen. In einer Lernphase, die sich normalerweise über die ersten zwei Seiten eines Textes erstreckt legt das Programm die fraglichen Stellen vor, die der Benutzer dann durch eine spezielle Codierung des Zehnerblocks entweder als korrekt bestätigt oder verbessert, indem er falsche Zuordnungen überschreibt, Fragmentiertes zusammensetzt, fälschlich Zusammengezogenes zerlegt und so weiter.

Das Programm merkt sich diese Korrekturen und baut sie in sein Regelwissen ein, es „lernt“ also aus seinen Fehlern. Die Fehlertoleranz kann der Benutzer regulieren, indem er den Wahrscheinlichkeitsgrad bestimmt der für die Einstufung als korrekt genügen soll.

Wie sieht es nun mit dem praktischen Nutzen eines solchen ICR-Systems aus? Neben der Kostensenkung im kommerziellen Bereich – etwa bei der Erfassung von Büchern für Neuauflagen oder der Archivierung von Geschäftsunterlagen ist diese Frage vor allem dort interessant, wo der Einsatz von Lesegeräten nicht nur eine Fortsetzung etablierter Praktiken mit anderen Mitteln bedeutet, sondern zugleich ein innovatives Umgehen mit Texten nach sich zieht. Nikolaus Tiling, Computerexperte am Literaturwissenschaftlichen Seminar der Universität Hamburg, schätzt die momentane Situation folgendermaßen ein:

„Die Akzeptanz gegenüber der maschinellen Texterkennung ist in unserem Bereich noch relativ gering. Das liegt einerseits an der Unzulänglichkeit der bisherigen Techniken, andererseits an einer immer noch verbreiteten Computerphobie. Immerhin gibt es ein Projekt, bei dem wir seit einiger Zeit erfolgreich OCR einsetzen: die elektronische Speicherung des Gesamtwerks von Hubert Fichte, das – unter Einschluß der zahlreichen Hörfunk-Skripte – rund 25 000 Seiten umfaßt. Das Ziel ist eine Volltextdatenbank, die sowohl für Motivrecherchen als auch für editorische Forschungen und die Auflösung von Datierungsfragen zur Verfügung stehen soll. ICR wäre für uns insofern interessant, als wir es selbst bei diesem modernen Autor häufig mit schwierigen

Druckvorlagen zu tun haben, ganz zu schweigen von Texten aus dem 18. Jahrhundert.

Selbstverständlich wird es auch in Zukunft nötig sein, eingeleseene Texte zu überarbeiten. Schon jetzt aber ist der Aufwand hiedür akzeptabel, zumal bei den neueren Systemen stumpfsinnige Arbeiten immer mehr wegfallen. Wenn sich der Korrekturleser dank einer guten Spalten- und Schriftstilerkennung kaum noch um Formatierungen kümmern muß oder sich dank Einzelblatteinzug und Multitasking unmittelbar an die Orthographie-prüfung machen kann, statt gelangweilt vor dem Scanner zu sitzen, dann bedarf es nicht erst der von hochwertigen ICRSystemen im Idealfall erreichbaren Fehlerquote von weniger als ein Promille, um diese Technik sinnvoll einzusetzen Insgesamt glaube ich, daß wir erst am Anfang einer stürmischen Entwicklung stehen, die dank sinkender Preise und wachsender Rechnerleistung die maschinelle Texterkennung zu einer Selbstverständlichkeit macht und in der Konsequenz unser Fach vorvöllig neue Aufgaben stellt."

Eine dieser künftigen Aufgaben wird wohl in der Vermeidung dessen bestehen, was in der kommerziellen Anwendung der OCR vorrangiges Ziel ist: das selektive Zugreifen auf Textbruchstücke, das den Vorgang des Lesens auf eine bloße Informationsbeschaffung reduziert Aber vielleicht wird uns ja eine ebenso stürmische Entwicklung der Sprachsynthesizer einmal jenes Behagen vermitteln, das schon Nietzsche von einer guten Vorlesemaschine erwartet haben dürfte: die CD mit dem Gesamtwerk des Lieblingsautors einlegen, den gewünschten Titel anwählen, zurücklehnen und den Gert-Westphal-Knopf ganz weit aufdrehen ...