LERNFÄHIGE COMPUTER

Texte automatisch zu übersetzen und aus grossen Datenbeständen Informationen herauszufiltern, ist die grosse Herausforderung der Computerlinguistik. Mit statistischen Methoden soll dieser Traum verwirklicht werden. Von Felix Würsten

«Peter schlug den Mann mit dem Stock.» Ein trügerisch einfacher Satz. Schlug Peter mit einem Stock einen Mann – oder schlug er einen Mann, der einen Stock bei sich hat? Was für den Menschen in der Regel einfach zu interpretieren ist, stellt Computer vor grosse Probleme. Denn wie in aller Welt soll die Maschine zuverlässig erkennen, was mit einem solch einfachen Satz gemeint ist? Genau mit dieser Grundfrage beschäftigen sich Michael Hess und sein Kollege Martin Volk, beide Professoren am Institut für Computerlinguistik der Universität Zürich. Ihr Ziel ist es, dem Computer das selbständige «Verstehen» von Texten beizubringen. Ein tückenreiches Unterfangen: «Unsere Sprache ist voller versteckter Mehrdeutigkeiten», bringt Hess das Problem auf den Punkt.

Ein erster naheliegender Ansatz ist, dem Computer die Regeln der Grammatik einzugeben und ihn dann die Texte gemäss diesen Regeln analysieren zu lassen. Bis zu einem gewissen Punkt ist dieser Ansatz auch erfolgreich. Doch gerade bei Mehrdeutigkeiten kommt man mit der Anwendung von Regeln schnell einmal nicht mehr weiter. Hess und Volk verfolgen daher auch statistische Ansätze. Der Computer lernt durch die Analyse von grossen Textmengen, wie Sätze zu verstehen sind und wie die einzelnen Satzteile zusammenhängen. Hilfreich ist vor allem, wenn der Computer übersetzte Texte vergleichen kann, ist doch eine mehrdeutige Aussage in der einen Sprache in der anderen oft eindeutig. Mit der Zeit «lernt» das Computerprogramm, dass die Kombination «Stock» und «schlagen» häufiger vorkommt als das Begriffspaar «Stock» und «Mann».

Eine konkrete Anwendung dieser Textanalyse erarbeitete Martin Volk in Zusammenarbeit mit einer skandinavischen Untertitelfirma. Sein Team entwickelte ein Programm, mit dem schwedische Filmuntertitel maschinell ins Dänische und Norwegische übersetzt werden können. Der Computer analysiert dabei grosse Mengen bereits übersetzter Untertitel auf charakteristische Wortfolgen und wiederkehrende Muster hin. Darauf basierend kann das Programm neue Texte übersetzen. «Die Software macht den Einsatz von Übersetzern nicht überflüssig, aber sie spart doch immerhin 20 Prozent Arbeitszeit», berichtet Volk.

TECHNISCHE HANDBÜCHER DURCHKÄMMEN

Von Interesse ist die neue Methode auch bei Textabfragen. Bei komplexen technischen Anlagen – Flugzeugen oder Atomkraftwerken beispielsweise – werden alle technischen Informationen in Handbüchern zusammengefasst, die oft mehrere zehntausend Seiten umfassen. Die Frage ist nun: Wie findet man in diesen umfangreichen Dokumenten schnell und zuverlässig eine Antwort auf eine konkrete Frage. «Eine Suche nach Stichworten findet zwar alle Seiten, auf denen die Begriffe auftauchen», erläutert Hess. «Doch ob diese Seiten die konkrete Frage beantworten, ist vorerst völlig unklar. Der Computer muss deshalb lernen zu erkennen, wann zwischen den Begriffen ein direkter inhaltlicher Zusammenhang besteht, damit er die gewünschten Informationen zuverlässig herausfiltern kann.»

Solche Abfragesysteme sind auch für Pharmafirmen interessant. Für die Forscher in den Entwicklungsabteilungen wäre es ein Fortschritt, wenn sie wissenschaftliche Publikationen systematisch absuchen könnten, ob darin Aussagen beispielsweise zu bestimmten Genen und Proteinen vorkommen. Ziel wäre es, aus den rund 70 Millionen Artikeln, die auf Medline zur Verfügung stehen, diejenigen Publikationen herauszusuchen, in denen die Aussage «A tut B in Bezug auf X» zu finden ist. Hess und seine Mitarbeiter sind daran, in einem gemeinsamen SNF-Projekt mit dem Pharmakonzern Novartis eine entsprechende Software zu entwickeln.

GIGANTISCHE TEXTMENGEN

Die statistische Textanalyse kommt auch bei einem Projekt zum Einsatz, das Volk gemeinsam mit Noah Bubenhofer vom Deutschen Seminar der Universität Zürich und dem Schweizer Alpen-Club bearbeitet. Die Wissenschaftler sind daran, alle Alpen-Jahrbücher von 1864 bis heute zu digitalisieren und auszuwerten. Beim Forschungsprojekt geht es darum, bestimmte Informationen aus diesen Jahrbüchern herauszufiltern, zum Beispiel welche Personen von 1880 bis 1920 eine Bergführerausbildung gemacht haben. «Das Programm muss Personennamen zuverlässig identifizieren und erkennen, dass in dieser Textpassage auch von Bergführern und Ausbildung die Rede ist», erklärt Volk.

Computerlinguisten seien heute gefragte Spezialisten, halten Hess und Volk fest. In vielen Bereichen stehen gigantische Textmengen zur Verfügung, die möglichst gezielt verarbeitet werden müssen. Die Palette von Anwendungen reicht dabei sehr weit: «Die Leute, die wir hier ausbilden, arbeiten später beispielsweise bei Normierungsgremien, die technische Texte möglichst effizient übersetzen oder Formulierungen auf mögliche Mehrdeutigkeiten hin untersuchen müssen», erzählt Hess. «Auch Grosskonzerne suchen nach Experten, die digitale Informationen systematisch durchforsten können, zum Beispiel um in der Berichterstattung im Internet Entwicklungen, die sich negativ auf das Unternehmen auswirken könnten, frühzeitig zu erkennen.»

«Unsere Sprache ist voll versteckter Mehrdeutigkeiten, für Computer ist sie deshalb schwierig zu interpretieren.» Michael Hess, Computerlinguist

KONTAKT Prof. Michael Hess, mhess@cl.uzh.ch; Prof. Martin Volk, volk@ifi.uzh.ch