

Von der Wissenschaft, das Internet zu befragen

Im Gespräch: Lichtenberg-Professorin Iryna Gurevych von der Universität Darmstadt über automatische Textanalyse im Internetzeitalter

Sie stehen noch am Beginn ihrer Karriere und haben doch längst ein neues Forschungsfeld für sich abgesteckt – die erfolgreichen Bewerberinnen und Bewerber um eine Lichtenberg-Professur der VolkswagenStiftung. Dieses Förderinstrument gibt ausgewählten exzellenten Forschern die nötige Freiheit und die Ausstattung, ihre hochinnovativen Ideen verfolgen zu können. Und an den Universitäten, die in den Genuss einer Lichtenberg-Professur kommen, lassen sich auf diesem Weg neue Forschungsgebiete etablieren.

Wer mit Iryna Gurevych ein Interview führt, muss sich in Acht nehmen: Die W1-Lichtenberg-Professorin ist Expertin für missverständliche Fragen und klare Antworten. An der Technischen Universität (TU) Darmstadt erforscht die 33-Jährige, wie Menschen aus der Informationsflut des Internets intelligente Antworten fischen können – und dabei die Frage nicht anders formulieren müssen, als sprächen sie mit ihrem Nachbarn. Iryna Gurevych versteht die Sprache von Mensch und Maschine. Sie hat Sprachwissenschaften studiert und forscht heute mit ihrer Lichtenberg-Professur – ausgestattet mit knapp 900.000 Euro – an einer Informatik-Fakultät. Mit ihr sprach Wissenschaftsjournalist *Frank van Bebber*.

Frau Professorin Gurevych, schauen Sie öfter ins Internet als in die Uni-Bibliothek?

Ich muss gestehen, ich war nur ein Mal in der Uni-Bibliothek – weil ich den Leiter der Bibliothek besucht habe, um mit ihm ein Entwicklungsprojekt zu virtuellen Forschungsumgebungen zu konzipieren. Ansonsten setze ich stark auf elektronische Wissensquellen und elektronische Bibliotheken. In der Informatik ändert sich das Wissen so schnell, dass wir kaum einschlägige Bücher haben; wir müssen uns auf das stützen, was wir in elektronischer Form im Internet finden.

Im Vergleich zur Uni-Bibliothek ist das Internet riesig. Ist es auch für Sie schwer, auf eine Frage eine Antwort zu bekommen?

Das Internet selbst gibt meist keine Antworten auf konkrete Fragen. Es gibt aber andere Wissensquellen im Web 2.0, zum Beispiel soziale Frage-Antwort-

Sie bringt dem Computer bei, ähnlich zu denken wie ein Mensch. Die Lichtenberg-Professorin Dr. Iryna Gurevych entwickelt an der Technischen Universität Darmstadt das sogenannte „Question Answering“, eine Technologie, mit der Computer Sinnzusammenhänge der menschlichen Sprache erfassen und natürlich-sprachliche Fragen „verstehen“ und richtig beantworten können.



Gibt man bei *Google* den Befehl „Kuchen - (minus) Obst“ ein, weil man auf der Suche nach Backrezepten OHNE Obst ist, antwortet die Suchmaschine mit Unmengen von Rezepten für Apfelkuchen und Kirschtorten. Diplom-informatiker Torsten Zesch, hier kurz vor dem Verzehr eines Kuchenstücks, weiß warum: *Google* versteht den Sinnzusammenhang zwischen einzelnen Fruchtarten und dem Oberbegriff „Obst“ nicht und sucht lediglich nach Seiten, auf denen der Begriff „Kuchen“ ohne das Wort „Obst“ auftaucht.

Plattformen. Bei diesen bekommt man zumeist sehr schnell Antworten von anderen Menschen, die auf dem einschlägigen Gebiet ausgewiesene Experten sind. Dieses Phänomen hat die Bezeichnung *Wisdom of Crowds*, auf Deutsch: die Weisheit der Vielen. Wir erforschen zum Beispiel Algorithmen: Wie etwa lassen sich Hunderte von Antworten auf eine Frage automatisch zusammenfassen? Das Motto unserer Arbeit könnte lauten: „Wenn Sie Fragen haben, kommen Sie nach Darmstadt!“

An welchen Fragetechniken für das Internet arbeiten Sie? Geben Sie bitte mal ein Beispiel!

Wir arbeiten an einer Technologie, die man unter der Bezeichnung *Question Answering* kennt: die Beantwortung natürlich-sprachlicher Fragen. Wir wollen intelligente Fragen stellen können und darauf intelligente Antworten erhalten. Das heißt: Wie finde ich in einem Dokument genau jene Stellen, die meine Frage beantworten – und nicht nur jene, in denen die Wörter der Frage vorkommen. Ein Beispiel: Sie wollen wissen, welche Professoren an der Technischen Universität Darmstadt tätig sind. Da würden Sie heute keine Antwort bekommen, sondern nur eine Liste mit Treffern, wo die Worte Professor und TU Darmstadt gleichzeitig auf einer Webseite auftauchen. Oder Sie interessiert, welche Vorteile Darmstadt als Informatik-Standort hat. Unsere Methoden ermöglichen es, Meinungen in Internetforen zu analysieren und zusammenzufassen.

Und das ist eine Aufgabe für eine Sprachwissenschaftlerin?

Ich habe tatsächlich mein Diplom in englischer und deutscher Linguistik gemacht. Mein akademischer Werdegang hat mich aber immer weiter weg von der Linguistik in Richtung Informatik geführt. Ich habe in Computerlinguistik promoviert – über Mensch-Maschine-Schnittstellen, die auf gesprochener Sprache beruhen. Anders gesagt: Das sind Informationssysteme, mit denen Sie reden können. Ich war dann fünf Jahre an einem Forschungsinstitut für angewandte Informatik tätig, bevor ich in den Fachbereich Informatik der TU Darmstadt gewechselt bin. Die Forschung auf dem Gebiet automatischer Textanalyse erfordert aber auch viele Kenntnisse über die Sprache. Es ist sehr komplex, dem Computer beizubringen, Zusammenhänge abzuleiten, die dem menschlichen Denken ähnlich sind. Der Computer muss den Sinn erschließen. Und wir möchten nun mit dem „Lichtenberg-Team“ Methoden entwickeln, die allgegenwärtig, benutzerfreundlich und flexibel sind – so wie Sie *Google* für Suchen aller Art benutzen können.

Nun gibt es im Internet auch viele Gerüchte und falsche Informationen. Wie soll ein Computer das erkennen?

Die Frage stellt sich gerade in Web 2.0-Kontexten, weil es hier keine Kontrolle zum Beispiel durch Verlage gibt. Ich habe kürzlich einen Doktoranden eingestellt, der sich mit automatischer Qualitätsbewertung von Webtexten befasst. Das ist schwer zu operationalisieren. Was ist die Qualität eines Textes? Manchmal kann eine Antwort willkommen sein, auch wenn sie voller Rechtschreibfehler ist. Doch viele Fehler sind manchmal auch Zeichen für Schlampeigkeit. Eine Vielzahl an automatisch bestimmbareren Qualitätsmerkmalen eines Textes erlaubt intelligentes Filtern. Es gibt im Internet aber auch Gemeinschaften, die Qualität sicherstellen: *Wikipedia* ist dadurch in vielen Bereichen die Informationsquelle Nummer eins.

Werden Computer einmal bessere Antworten im Internet aufstöbern können als Menschen?

Ich würde sogar so weit gehen zu sagen, dass das heute schon der Fall ist. In vielen Bereichen gibt es derart viele Texte, dass niemand die mehr alle lesen kann. In begrenzten Bereichen ist natürlich der menschliche Experte weiterhin die Lösung Nummer eins.

Wie arbeitet Ihr Team an den neuen Methoden, Informationen aus dem Internet zu heben?

In meinem Team sind 17 wissenschaftliche Mitarbeiterinnen und Mitarbeiter beschäftigt. Wir arbeiten in vier Gruppen mit zentral jeweils einem Post-Doktoranden zu einer Leitfrage, die bisher nicht erforscht ist. Zum Beispiel ent-

Das 17-köpfige Forscherteam um Lichtenberg-Professorin Dr. Iryna Gurevych (vordere Reihe, Zweite von links) auf dem Campus der TU Darmstadt. Dank fünfjähriger Lichtenberg-Professur konnte Iryna Gurevych ein völlig neues Forschungsgebiet definieren und aufbauen. Und auch langfristig bieten sich für das Team gute Perspektiven an der Universität, da sich das außerordentliche Potenzial des Forschungsfeldes bereits bestätigt hat.



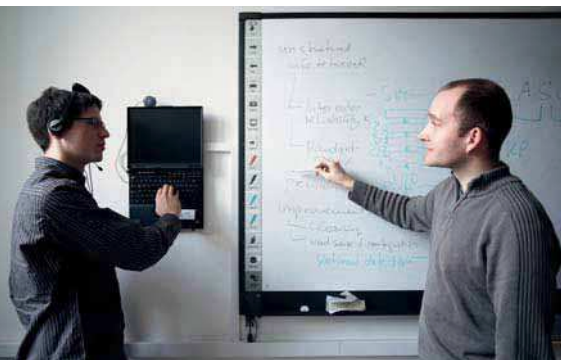
wickeln wir Algorithmen, die die Ähnlichkeit zwischen zwei Texten automatisch bestimmen. Dazu müssen wir erst einmal herausfinden: Was macht Texte zu ähnlichen Texten? Eine Ähnlichkeit im Sinne des Themas? Oder sind es identische Wörter? Dann überlegen wir uns, wie wir dies in einen Algorithmus verwandeln. Am Ende prüfen wir die Ergebnisse der Software anhand einer von Menschen erstellten Musterlösung. Wir arbeiten mit Unternehmen zusammen; vieles aber ist Grundlagenforschung, die ja Gegenstand der Lichtenberg-Professur ist. Beides in Kombination macht dies von der Stiftung geförderte Vorhaben so spannend.

Wie hat Ihnen die Lichtenberg-Professur geholfen, mit Geld?

Es ist sicher nicht das Geld. Ich habe viele andere Drittmittel-Forschungsprojekte. An der Lichtenberg-Professur hat mich fasziniert, dass ich ein originäres Forschungsgebiet definieren und international aufbauen kann. Das ist ein Risiko, aber auch ein Reiz, weil man der Erste ist und die Chance hat, ein Forschungsgebiet mit seinem Namen zu verbinden. Und ich habe eine langfristige Perspektive an der Universität, weil bei Lichtenberg-Professuren gleich eine Zielvereinbarung über ein *Tenure-track*-Verfahren geschlossen wird. Man weiß: Nach fünf Jahren ist es nicht vorbei, wenn sich das Potenzial bestätigt. Eine weitere Motivation für mich war, dass Lichtenberg-Professuren interdisziplinär angelegt sind. Ich bin ja durch verschiedene wissenschaftliche Disziplinen geprägt, und die Professur macht es möglich, diese in einem Forschungsgebiet zu verbinden. Das Potenzial meiner Forschung hat sich bereits bestätigt – mit Rufen auf gleich zwei W3-Professuren.

Als Sie die Professur antraten, sind Sie gerade 32 Jahre alt geworden. Weckt da der Anspruch, ein neues Forschungsfeld eröffnen zu wollen, nicht Zweifel bei sich und anderen?

Ich versuche, mit meiner Arbeit zu überzeugen. Mir ist der Einstieg aus zwei Gründen leichter gefallen: Zum einen habe ich einen Schritt nach dem anderen gemacht. An die Technische Universität Darmstadt kam ich Ende 2005 zunächst als leitende wissenschaftliche Mitarbeiterin im Forschungsschwerpunkt *E-Learning* und als Leiterin eines von mir eingeworbenen Projekts bei der Deutschen Forschungsgemeinschaft. Nach anderthalb Jahren kam eine Emmy-Noether-Gruppe der DFG hinzu, dann die Lichtenberg-Professur. Der andere Punkt ist, dass ich mit meinen Kolleginnen und Kollegen wunderbar zusammenarbeiten kann; ich habe eine exzellente Arbeitsgruppe aufgebaut, die mir viel bedeutet. Gemeinsam ist es uns inzwischen gelungen, den Standort TU Darmstadt auf dem Gebiet der automatischen Textanalyse weltweit bekannt zu machen. Wir haben manches entwickelt, was auch in Stanford oder Berkeley eingesetzt wird. Mehr kann man in so kurzer Zeit wohl kaum erreichen.



Brainstorming im Demonstrations- und Teambesprechungsraum an der TU Darmstadt. Torsten Zesch, Cigdem Toprak, Dr. Delphine Bernhard, Professorin Iryna Gurevych und Christof Müller (Bild oben, von links) diskutieren über die Internetsuche der nächsten Generation. Das von den Forschern entwickelte „*Question Answering*“ könnte *Google* und *Co.* revolutionieren, indem es die Flexibilität und Effektivität von Suchmaschinen deutlich erhöht.

Warum sind Sie Ihren Entwicklungen nicht in die USA gefolgt?

Weil ich gern in Deutschland bin. Deutschland ist für mich ja schon Ausland. Ich komme aus der Ukraine, dort habe ich studiert. Ich war Stipendiatin des Deutschen Akademischen Austauschdienstes und habe hier promoviert. Ich lebe sehr gern in Deutschland, hier ist meine achtjährige Tochter geboren und wächst glücklich auf. Es stimmt: In den USA wird viel mehr im Bereich automatischer Textanalyse gearbeitet. Das ist zweifellos attraktiv, aber mit der Lichtenberg-Professur wurde mir ein traumhaftes Angebot unterbreitet – ganz einfach ein tolles Förderpaket geschnürt. Von den Kollegen in den USA werde ich mittlerweile oft als Gastsprecherin und Partner für Forschungsprojekte eingeladen.

Und wer das von Ihnen und Ihrem Team produzierte Wissen sucht, sollte im Internet schauen, nicht in der Bibliothek?

Ja. Sobald ein Aufsatz für eine Veröffentlichung angenommen wurde, stellen wir ihn auf unsere Webseite – ebenso wie Informationen über unsere Forschungsprojekte. Unsere *Community* pflegt die *Open-Access*-Kultur. Die Leute finden die Informationen von und über uns entweder mit *Google* oder sie schauen gezielt auf unserer Webseite danach.

Frau Gurevych, vielen Dank für das Gespräch.

Die Lichtenberg-Professuren

Exzellente Forscherinnen und Forscher benötigen zweierlei: die Freiheit, eigene Ideen zu verfolgen, und eine Ausstattung, mit der die Realisierung dieser Ideen möglich wird. Beides erhält, wer sich bei der VolkswagenStiftung erfolgreich um eine Lichtenberg-Professur bemüht. Zugleich können Universitäten mithilfe einer solchen Professur auf besonders substanzielle und nachhaltige Weise Strukturplanung betreiben; das macht dieses Förderinstrument für mehr und mehr Hochschulen attraktiv. Fünf bis acht Jahre lang unterstützt die Stiftung die herausragenden Wissenschaftlerinnen und Wissenschaftler dabei, ihr Forschungs-

feld an einer Universität ihrer Wahl fest zu verankern. 25 solcher Professuren – benannt nach dem Mathematiker, Physiker und Philosophen Georg Christoph Lichtenberg – hat die Stiftung zwischen 2003 und 2009 an 17 deutschen Hochschulen eingerichtet (im Sommer 2009 sind vier weitere hinzugekommen, die ihre Arbeit allerdings noch nicht aufgenommen haben). All diese Wissenschaftlerpersönlichkeiten eint, dass sie frische Strategien in das traditionelle Hochschulsystem einspeisen. Auf diese Weise gelingt es, bestehende Strukturen aufzubrechen und Neuem nachhaltigen Weg zu bahnen.

cj