Wie Computerlinguisten arbeiten / Forscher lesen mit Hilfe des Computers riesige Textmengen

Wie Computerlinguisten arbeiten / Forscher lesen mit Hilfe des Computers riesige Textmengen
Der digitale Wind weht quer durch alle textbasierten Geisteswissenschaften, von Literatur bis Soziologie. Wie gehen Forscher mit den Unmengen an Daten um? Wissenschaftler der Universitäten Stuttgart, Hildesheim und Potsdam nehmen sich nun Zeitungsarchive vor. Dabei untersuchen Politikwissenschaftler, welche kollektiven Identitäten - etwa europäische, nationale, religiöse - im Zusammenhang mit internationalen Krisen genannt werden.
Die Computerlinguisten der Hildesheimer Universität sind am Anfang der Kette. "Wir bringen Ordnung in journalistische Textarchive", sagt Professor Ulrich Heid. Statt einer Einzellektüre gehen sie eine große Datensammlung durch und suchen nach Mustern. Politikwissenschaftler analysieren Texte bisher meist manuell - aufwändig und punktuell - oder mit bestehenden Werkzeugen, mit denen sie aber nur einige Tausend Artikel bearbeiten oder nach Wortformen suchen können. Eine tiefergehende sprachbezogene Analyse ist nicht möglich. Unterstützt durch computerlinguistische Verfahren sollen die Fachwissenschaftler nun große Mengen eigenständig bearbeiten können. So können sie zum Beispiel sehr schnell entscheiden, ob ein Artikel zum Thema "Krisen, Krieg, militärische Interventionen" gehört oder nicht. Denn auch in Fußballberichten findet sich viel einschlägiges Vokabular: Da wird geschossen, verteidigt, eine Linie gehalten.
Aber wie findet man "Identität" in riesigen Textmengen? "Wir analysieren das Umfeld, nicht einzelne Worte, wir suchen nach Mustern, etwa Formulierungen wie ?x zeigte sich erfreut. Dann rechnen wir zurück, wer spricht, worüber spricht er, wertend oder nicht wertend", sagt Ulrich Heid. Kollektive Identitäten können ganz unterschiedlich ausgedrückt werden, in journalistischen Texten sind oft Andeutungen und Metaphern enthalten. Typisch sind etwa Ausdrücke wie "Washington kann in dieser Frage nicht über seinen Schatten springen". "Wir suchen auch versteckte Informationen in Texten. Man muss hinter die Formulierung schauen und tiefer in den Text einsteigen." Heid nennt ein weiteres Beispiel: Wenn die Bundeskanzlerin von "wir" spricht oder ein Politiker "Wir haben gewonnen" sagt, so kann dahinter viel stecken: "Wir" kann die Partei, Europa, das Land oder eine niedersächsische Provinz meinen. Daher betrachten die Linguisten so etwas Spezifisches wie "wir" im Kontext.
Zunächst sammeln die Forscher, welche Zeitungen über Kriege und humanitäre Interventionen seit 1990 geschrieben haben. Sie greifen auf etwa 800.000 Zeitungsartikel europäischer Länder - Österreich, Deutschland, Irland, Frankreich, Großbritannien - und der USA zurück (Januar 1990 bis Dezember 2012), darunter die Frankfurter Allgemeine Zeitung, Süddeutsche Zeitung, Le Monde, The Guardian und die New York Times, unter Lizenz von kommerziellen Zeitungsarchiven. Dabei unterscheiden die Forscher zwischen Kommentaren, Meldungen, Leitartikeln und weiteren Textarten, markieren ähnliche Artikel und Dubletten von Nachrichtenagenturen und finden Wege, Fußballberichte mit "Kriegsterminologie" auszusortieren.
"Wir haben es mit digitalen Daten in unterschiedlichen Formaten und Datenstrukturen zu tun. Datenmaterial aus verschiedenen Quellen einheitlich aufzubereiten ist komplex", sagt Fritz Kliche, wissenschaftlicher Mitarbeiter am Institut für Informationswissenschaft und Sprachtechnologie der Uni Hildesheim. Dabei haben die Forscher Erfahrung in der Zeitungsanalyse: So hat Ulrich Heids frühere Arbeitsgruppe in Zusammenarbeit mit dem Max-Planck-Institut für internationales Strafrecht etwa eine halbe Millionen Artikel zum Thema Familientragödien analysiert - auf der Suche nach Tatmustern.
Fachwissenschaftler können die Texte - je nach Forschungsfrage - nach Wörtern und Wortsequenzen durchsuchen oder nach einer großen Anzahl von inhaltlich ähnlichen oder sprachlich unterschiedlichen "Sprechweisen". Die Stuttgarter Politikwissenschaftlerin Professorin Cathleen Kantner, die das Verbundprojekt leitet, hat eine Vielzahl von Sprechweisen identifiziert, die auf einen Bezug auf Europa als Wertegemeinschaft hindeuten. Besonders ist dabei, dass für alle Belege der Publikationszeitpunkt und andere Metadaten bekannt sind: Rückblickend kann man darstellen, was eine Ankündigung auslöst - etwa die Energiewende nach dem Unglück in Fukushima - oder wie sich die Einstellungen zum "arabischen Frühling" verändert haben. Die Medienaufmerksamkeit für ein Thema kann somit weitgehend automatisch errechnet und in einer Grafik als Kurve über Tage, Wochen oder Monate dargestellt werden. Darauf können dann Detailuntersuchungen aufsetzen.
Das Bundesforschungsministerium fördert das dreijährige Projekt "eIdentity" bis 2015 mit insgesamt 853.000 Euro. Das Verbundprojekt wird von der Universität Stuttgart koordiniert. Derzeit können die computerlinguistischen Verfahren auf Texte in deutscher, englischer und französischer Sprache angewandt werden.
Beispiele:
Wer in der Ukraine-Krise an die europäische Identität appelliert
"Bei einer Tagung in Berlin, im Beisein von Bundeskanzlerin Angela Merkel, erinnerte Barroso daran, wie viel friedlicher Wandel schon mit Hilfe der europäischen Ideale erreicht worden sei, vom Mauerfall in Berlin bis zum Ende der Diktatur in seinem Heimatland Portugal - und Ähnliches müsse nun auch in der Ukraine geschehen, gerade nach der jüngsten russischen Intervention auf der Krim."
(Ukraine-Russland-Krise: Das Dilemma der Europäer, SPIEGEL online, 01.03.2014, www.spiegel.de/politik/ausland/ukraine-krim-eu-das-dilemma-der-europaeer-a-956438.html)
"Wenn wir Europäer mit Europäern in der Ukraine reden, ist das keine Einmischung in innere Angelegenheit, sondern eine Selbstverständlichkeit."
(Westerwelle warnt Ukraine vor Gewalt gegen Opposition, ZEIT online, 05.12.2013, www.zeit.de/news/2013-12/05/eu-westerwelle-warnt-ukraine-vor-gewalt-gegen-opposition-05140011)
"Es war gleich zu Beginn des Konflikts um die Ukraine eine Schwäche der westlichen Europäer, dass sie nichts eilfertiger ausschlossen als ?militärische Optionen. Auch die große Koalition in Berlin einigte sich sehr schnell auf diese Sprachregelung und schwang sich damit auf die westeuropäische Wolke der Ängstlichkeit [...]."
(Krise in der Ukraine: Putins Optionen, Frankfurter Allgemeine Zeitung, online, 21.04.2014, www.faz.net/aktuell/politik/ausland/europa/krise-in-der-ukraine-putins-optionen-12903959.html)
Worüber Computerlinguisten forschen - ein weiteres Beispiel:
DFG fördert Forschungsprojekt: Pfiffig, intelligent, klug
"Sentiment Analysis" - das ist die automatische Bestimmung von Meinungen in Texten. Wird eine Person, eine Sache oder ein Sachverhalt neutral geschildert oder zusätzlich positiv oder negativ bewertet? Wie liegen Bewertungen auf einer Intensitäts-Skala: was ist stärker, was weniger positiv: pfiffig, intelligent, klug? Um das bestimmen zu können, arbeitet Dr. Josef Ruppenhofer in einem von der Deutschen Forschungsgemeinschaft geförderten Projekt an sprachtechnologischen Methoden und Computerverfahren für Englisch und Deutsch. Wörter alleine genügen oft nicht, um die Bewertung zu bestimmen braucht man auch den Kontext: warmes Badezimmer gilt als positiv, warmes Bier meistens nicht.
Für Deutsch gibt es noch wenige Hilfsmittel, und das Interesse ist groß, weil zum Beispiel viele Firmen "Opinion Mining" betreiben: sie wollen wissen, wie sich die Menschen in den sozialen Medien über die Firma und ihre Produkte äußern - dafür werden die Ergebnisse des Projekts eine solide Grundlage sein. Josef Ruppenhofer bereitet in diesem Bereich seine Habilitation am Institut für Informationswissenschaft und Sprachtechnologie der Uni Hildesheim vor und kooperiert im Projekt mit Kollegen an der Universität des Saarlandes.
Konferenz "Sprachtechnologie und Computerlinguistik":
Vom 8. bis 10. Oktober 2014 richtet die Universität Hildesheim die 12. Tagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL, der wissenschaftliche Fachverband für maschinelle Sprachverarbeitung) und der Österreichischen Gesellschaft für Artificial Intelligence (ÖGAI, Künstliche Intelligenz) aus. An der KONVENS-Konferenz ist auch eine Arbeitsgruppe der Deutschen Gesellschaft für Sprachwissenschaft (DGfS) beteiligt. Etwa 100 Fachleute aus europäischen Ländern tauschen sich auf hohem Niveau zu computerlinguistischer Grundlagenforschung aus. Konferenzsprachen sind Englisch und Deutsch.
Programm: www.uni-hildesheim.de/konvens2014/

Medienkontakt:
Pressestelle,
Isa Lange
E-Mail: presse@uni-hildesheim.de
Telefon: 05121. 883-90100 und 0177.860.5905
<img src="http://www.pressrelations.de/new/pmcounter.cfm?n_pinr_=577152" width="1" height="1">

Pressekontakt

Stiftung Universität Hildesheim

31141 Hildesheim

presse@uni-hildesheim.de

Firmenkontakt

Stiftung Universität Hildesheim

31141 Hildesheim

presse@uni-hildesheim.de

Weitere Informationen finden sich auf unserer Homepage