idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Carl von Ossietzky-Universität Oldenburg
Dr. Corinna Dahm-Brey

19.09.2014
http://idw-online.de/de/news604290

Wissenschaftliche Publikationen
Informationstechnik
überregional

Computer als Waschmaschine für verunreinigte Texte - Entwickelte Software "kann" sogar klingonisch

Eine Methode zur automatischen Reinigung verschmutzter und bekritzelter Texte etwa von Kaffeeflecken oder Durchstreichungen haben Forscher der Universitäten Oldenburg, Frankfurt am Main, Sheffield (Großbritannien) und der Technischen Universität Berlin entwickelt. Wie sich ein Computer nebst Scanner und Drucker mithin als „Waschmaschine" für Texte einsetzen lässt, veröffentlicht der interdisziplinäre Oldenburger Forscher Prof. Dr. Jörg Lücke gemeinsam mit seinem Sheffielder Kollegen Dr. Zhenwen Dai in der Oktober-Ausgabe der renommierten Fachzeitschrift TPAMI („IEEE Transactions on Pattern Analysis and Machine Intelligence").

Die neu entwickelte Software zur Textreinigung ist Ergebnis eines von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts. Für die Arbeiten unter dem Titel „Nicht-lineare probabilistische Modelle für repräsentations-basiertes Erkennen und unüberwachtes Lernen auf visuellen Daten" sind bisher etwa eine halbe Million Euro an Fördergeldern zugesagt.
Der Schlüssel zum Reinigungserfolg ist Statistik. Buchstaben – etwa in einem Zeitungsartikel – sind regelmäßige, sich wiederholende Muster, während Schmutz-Muster wie Kaffee- oder Tintenflecken sehr selten gleich aussehen. Das neu entwickelte Computerprogramm schaut sich einen verunreinigten Text zunächst viele Male an und lernt dabei, aus welchen sich regelmäßig wiederholenden Mustern (also Buchstaben) er besteht. Danach merkt sich das Programm die „saubersten" Beispiele für jeden Buchstaben, um Schritt für Schritt jeden einzelnen damit zu ersetzen. Das Ergebnis ist ein sauberer Text.
Besonderer Clou ist die Unabhängigkeit von Sprache oder Alphabet des Textes: Da das Programm zunächst die Buchstaben lernt, funktioniert es zum Beispiel auch mit einem Text in der Phantasiesprache Klingonisch (aus der Serie „Raumschiff Enterprise"). Ein weiterer Unterschied zu handelsüblichen Texterkennungs-Programmen ist seine Fähigkeit, mit besonders schweren Verschmutzungen umgehen zu können.
Eine Herausforderung stellt dabei bislang noch der große Bedarf an Rechenkapazität dar, wie Projektleiter Lücke berichtet: „Wegen des enormen Rechenaufwandes können wir derzeit nur recht kleine Alphabete behandeln, und dennoch benötigen wir einen Rechen-Cluster mit 15 Grafikkarten-Prozessoren, um zu den vorgestellten Ergebnissen zu gelangen." Eine direkte Anwendbarkeit sei aber auch nicht das primäre Ziel der Forschung gewesen, sondern zunächst die grundsätzliche Erprobung der neuen Methode. Von ihr könnten in Zukunft automatische Texterkennungs-Programme oder Software zur Restauration alter Zeitschriftentexte profitieren. Lücke sieht auch einen Nutzen der Resultate für die Erkennung gesprochener Sprache und die Analyse medizinischer Bild-Daten. „In beiden Fällen stellen starke ‚Verschmutzungen' in der Form von Rauschen und Signal-Verzerrungen derzeit die größten Herausforderungen dar." Ein Beispiel dafür seien die oft schlechten Leistungen heutiger Spracherkennungs-Programme bei Hintergrundgeräuschen. „Mit unserer neuen Methode haben wir nun ein Werkzeug in der Hand, um diese Herausforderungen angehen zu können."

Zhenwen Dai and Jörg Lücke (2014): Autonomous Document Cleaning – A Generative Approach to Reconstruct Strongly Corrupted Scanned Texts. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 36(10): 1950-1962, 2004.

Kontakt: Prof. Dr. Jörg Lücke, Arbeitsgruppe Machine Learning und Exzellenzcluster Hearing4all, Tel.: 0441/798-3252 (Sekretariat), E-Mail: joerg.luecke@uni-oldenburg.de

URL zur Pressemitteilung: http://www.uni-oldenburg.de/ml - Arbeitsgruppe "Machine Learning" an der Universität Oldenburg

URL zur Pressemitteilung: http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=6777544 - Aufsatz auf IEEE-Website