Medieninformation

TECHNISCHE UNIVERSITÄT DARMSTADT

Für starke Argumente

TU Darmstadt: Neue Tools analysieren die Qualität von Texten im Internet

Kommunikation und Medien
Corporate Communications

Karolinenplatz 5
64289 Darmstadt

Ihr Ansprechpartner:
Jörg Feuck
Tel. 06151 16 - 20018
Fax 06151 16 - 23750
feuck@pvw.tu-darmstadt.de

www.tu-darmstadt.de/presse
presse@tu-darmstadt.de

Darmstadt, 19. Dezember 2016. Das Internet bietet eine Flut von Informationen und Argumenten zu vielen weltbewegenden Themen. Aber wie zuverlässig und glaubwürdig sind diese Texte und Debattenbeiträge von Experten und interessierten Laien? Software-Instrumente, die nicht nur automatisiert Pro- und Contra-Argumente aus Texten herausfiltern, sondern sie auch einem Qualitätstest unterziehen, sind bereits in der Entwicklung. Das Ubiquitous Knowledge Processing (UPK) Lab der TU Darmstadt stellt der Wissenschafts-Community jetzt eine erfolgversprechende Trainingsdatenbank zur Verfügung um neue Methoden zu testen.

„Argumente und Informationen aus dem Netz sind bislang weitgehend unvalidiert“, erklärt die Leiterin des UKP-Labs, Professor Iryna Gurevych. Sie und ihr Forschungsteam entwickeln Werkzeuge, die für unterschiedliche Anwendungsbereiche große Mengen an Informationen aus unterschiedlichen Kanälen erschließen und für den Nutzer vorstrukturieren. Mit der neuen Datenbank „UKPConvArg2“ haben sie jetzt ein Corpus geschaffen, das insgesamt 9.111 für die maschinelle Anwendung codierte Argumenten-Paare umfasst. Die Daten stammen aus 16 Social-Media-Debatten zu gesellschaftlich relevanten Themen. Rund 800 Crowdworker haben das Material auf der Basis von 17 Qualitätskriterien bewertet, Fachexperten diese Bewertung anschließend evaluiert.

Die Trainingsdatenbank, die der Wissenschafts-Community seit November zur Verfügung steht, zeigt nicht nur, welche Argumente überzeugend sind und warum. Sie bildet auch die Ausgangsbasis zur Entwicklung neuer Methoden für die empirische Analyse von Textdaten aus dem Internet. „Damit können wir eine neue Diskussion um die Möglichkeiten des maschinellen Lernens eröffnen“, betont Ivan Habernal, Wissenschaftler am UKP-Lab. Erste Experimente mit verschiedenen mathematischen Modellen zur Auswertung der Trainingsdaten seien bereits erfolgreich verlaufen.

Einfache Anwendungen wie die Segmentierung von Texten in Argumente, Fakten und Behauptungen innerhalb einheitlicher und umgrenzter Textsorten sind in naher Zukunft bereits umsetzbar. Eine große Herausforderung ist allerdings noch immer die Analyse vielfältiger, heterogener und komplexer Daten von unterschiedlicher Qualität von Fachaufsätzen bis hin zu Social-Media-Beiträgen wie sie zum Beispiel in der geistes- und sozialwissenschaftlichen Forschung gebraucht wird. Zum einen ist hier die Erstellung von Trainingsdaten hoch komplex, zum anderen können Methoden, die für eine bestimmte Textsorte entwickelt werden, bislang kaum auf andere

übertragen werden. „Die Skalierungsfrage haben wir noch nicht gelöst“, sagt Iryna Gurevych. „Das ist eine Forschungsaufgabe, der wir uns jetzt vordringlich widmen.“

Internet

Die neue Trainingsdatenbank „UKPConvArg2“ steht unter der Lizenz CC-BY-SA für weitere Experimente unter https://github.com/UKPLab/emnlp2016-empirical-convincingness zur Verfügung.
Weitere Informationen zum UKP-Lab: www.ukp.tu-darmstadt.de.
Mit den Forschungen am UKP-Lab befasst sich ausführlich ein Beitrag in der soeben erschienenen Dezember-Ausgabe der hoch³Forschen: http://www.tu-darmstadt.de/hoch3-forschen

Hintergrund: Future Data Analytics for Humanities

An der TU Darmstadt vernetzen und koordinieren sich Wissenschaftlerinnen und Wissenschaftler immer intensiver, um gemeinsam neue Methoden der Datenanalyse für die Geistes- und Sozialwissenschaften zu erforschen und weiterzuentwickeln. Bisherige Ansätze zur Analyse von geistes- und sozialwissenschaftlichen Daten scheitern oft an der Vielfalt, Komplexität und Heterogenität sowohl der Daten als auch der Fragestellungen. Die neuen Methoden müssen mit wenigen Trainingsdaten auskommen, mit Daten unterschiedlicher Qualität und Beschaffenheit umgehen können, kontinuierlich aus der Interaktion mit Forschenden lernen, Objekte aus verschiedenen Datenquellen semantisch analysieren und verknüpfen, und das Wissen aus externen Quellen auf gegebene und neue Fragestellungen übertragen.
Die Datenanalyse umfasst Forschungsrichtungen wie die automatische Sprachverarbeitung, Visual Computing sowie maschinelles Lernen. Prototypische Anwender in den Humanities sind Forschende aus Philosophie, Philologie, Geschichte, Online- und Kommunikationswissenschaft sowie der Archäologie.

MI-Nr. 90/2016, Witte/feu