



## Das Gleiche in Grün

Software der TU Darmstadt erkennt Texte, die mit anderen Worten etwas Ähnliches aussagen

Darmstadt, 20.06.2012. Eine an der TU Darmstadt entwickelte Software kann die Bedeutungen von Texten verstehen und erkennt zuverlässig ähnliche Aussagen – auch, wenn sie in anderen Worten ausgedrückt werden.

Für Computer sind freie Texte im Regelfall nicht mehr als eine Sequenz bestimmter Zeichen. Die Bedeutung dieser Zeichenketten, deren „Semantik“, bleibt dem Computer zunächst verborgen. Daher können Computer bislang auch nicht erkennen, wenn zwei Texte den gleichen Sachverhalt mit anderen Worten beschreiben. Um Rechnern beizubringen, dass etwa bei einem Hochwasser ein »sinkender Wasserstand« die gleiche Bedeutung hat wie die Aussage, der Pegel habe »seinen Scheitelpunkt überschritten«, haben Informatiker des Ubiquitous Knowledge Processing (UKP) Lab an der TU Darmstadt eine Software entwickelt und diese mit Hilfe eines maschinellen Lernverfahrens trainiert.

Anhand konkreter Texteigenschaften erschließt die Software die Bedeutung eines Textes und bestimmt aus einer Fülle von möglichen Indikatoren automatisch semantisch ähnliche Texte – auch wenn diese sich auf der Ebene der Zeichenketten, also der konkreten Worte, stark unterscheiden. Dazu nutzen die Wissenschaftler unter anderem maschinenlesbare Wörterbücher, sogenannte „Wortnetze“ sowie komplexe Modelle, die Texte in einem hochdimensionalen mathematischen Vektorraum repräsentieren und vergleichen. „Unser System funktioniert nicht nur auf bestimmten Daten, sondern ist mittlerweile auch auf vorab unbekanntem Daten sehr solide einsetzbar“, sagt die Leiterin des UKP Lab, Prof. Dr. Iryna Gurevych. Beim internationalen Wettbewerb Semantic Evaluation 2012 erzielte die Software bei der Aufgabe „Bestimmung semantischer Ähnlichkeit von Texten“ unter 35 konkurrierenden Teams die beste Leistung.

Die Software, die von Daniel Bär, Prof. Dr. Chris Biemann, Prof. Dr. Iryna Gurevych und Dr. Torsten Zesch entwickelt wurde, bietet zahlreiche Anwendungsmöglichkeiten: Unter anderem könnten Leser Empfehlungen für thematisch verwandte Dokumente erhalten und die Freitextantworten in groß angelegten Umfragen oder internationalen Kompetenzmessungen wie der PISA-Studie mit deutlich geringerem Aufwand ausgewertet werden.

Das UKP Lab ist in die internationale Forschung zu intelligenten Sprachtechnologien und Text-Mining eingebunden. Es arbeitet im Rahmen

Kommunikation und Medien  
Corporate Communications

Karolinenplatz 5  
64289 Darmstadt

Ihr Ansprechpartner:  
Christian Siemens  
Tel. 06151 16 - 32 29  
Fax 06151 16 - 41 28  
[siemens.ch@pvw.tu-darmstadt.de](mailto:siemens.ch@pvw.tu-darmstadt.de)

[www.tu-darmstadt.de/presse](http://www.tu-darmstadt.de/presse)  
[presse@tu-darmstadt.de](mailto:presse@tu-darmstadt.de)



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

der seit 2012 begonnenen institutionellen Kooperation zwischen der TU Darmstadt und dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt sowohl an anwendungsbezogener Theorie und Grundlagenforschung an der TU Darmstadt als auch an theoretisch fundierten Anwendungen der Sprachtechnologie in bildungsbezogenen Kontexten.

**Pressekontakt**

Prof. Dr. Iryna Gurevych

Tel. 06151/16-5411

[gurevych@ukp.informatik.tu-darmstadt.de](mailto:gurevych@ukp.informatik.tu-darmstadt.de)

**Weitere Informationen**

[www.ukp.tu-darmstadt.de](http://www.ukp.tu-darmstadt.de)

MI-Nr. 50/2012, csi