

Lingenio mit neuen Algorithmen auf der EAMT

Prof. Babych von Lingenio und der Universität Leeds stellt neues Verfahren für die MÜ-Datengewinnung vor

Heidelberg, 22. Juni 2018. Bogdan Babych, neuer Technischer Leiter bei der Lingenio GmbH und Professor an der Universität Leeds, trägt bei der diesjährigen Konferenz der [European Association for Machine Translation](#) (EAMT) in Alicante über eine neues Verfahren zur Gewinnung von Übersetzungsdaten aus sog. vergleichbaren Corpora mit großem Potenzial für die mehrsprachige Termextraktion vor.

Die EAMT veranstaltet jährlich eine Konferenz zu neuesten Entwicklungen in der Maschinellen Übersetzung. Dieses Jahr fand die Veranstaltung in Alicante statt. ([The 21st Annual Conference of the European Association for Machine Translation](#)). Ein wichtiges aktuelles Thema ist die Gewinnung von Übersetzungsdaten aus sog. 'vergleichbaren Corpora' ([Comparable Corpora](#)), d.h. von Texten unterschiedlicher Sprache die nicht Übersetzungen voneinander sind, sondern nur in der Thematik ähnlich.



Das ist gerade für Übersetzungen von und in sog. 'kleinere Sprachen' wichtig. Das sind solche, für die es vergleichsweise wenig elektronisch verfügbare Texte gibt und – vor allem – wenig Übersetzungen in 'große Sprachen' wie Englisch oder Französisch. Um Übersetzungssysteme trainieren zu können, muss dann auf andere, nur vergleichbare, Korpora ausgewichen werden. Für diese ist es von großer Bedeutung, sog. [Cognates](#) zu finden, d.h. Wörter in den Texten die mit großer Sicherheit als Übersetzung voneinander erkannt werden können. Bei Vorliegen von nur bescheidenen Wörterbüchern, wie oft z.B. zwischen asiatischen und europäischen Sprachen der Fall, ist das ein großes Problem.

Prof. Babych hat nach detaillierten Studien zu diesem Thema nun eine vielversprechende Arbeit vorgelegt die zur Bestimmung von Cognates zusätzliche, standardisierte und automatisch abfragbare Kriterien aus der Phonologie mit einbezieht. Die Arbeit ist auf großes Interesse bei Kollegen gestoßen, u.a. von der [School of Computing](#) der University of Dublin und der [Universität d'Alicant](#), beides bekannte Zentren innovativer Forschung zur automatischen Übersetzung. Lingenio wird das Verfahren in seine Term-Extraktions-Systeme zur Unterstützung von diversen Übersetzungsprozessen integrieren.

Mehr zu der Arbeit zu '*Development and evaluation of phonological models for cognate identification*' findet sich in den [Proceedings der Konferenz](#).