Sprachen sind nicht bloß Folgen von Wörtern

Online-Übersetzungssysteme lernen Sprachen, indem sie große Mengen von mehrsprachigen Texten mit statistischen Lernverfahren verarbeiten. Dabei berechnen sie die Wahrscheinlichkeiten bestimmter Wortfolgen. Sprachen sind aber nicht bloß Folgen von Wörtern, sondern haben eine grammatische Struktur. Das Language Technology Lab des Deutschen Forschungszentrums für Künstliche Intelligenz entwickelt zur Zeit eine hybride Übersetzungstechnologie, die statistische Systeme mit linguistischem Wissen und semantischem Wissen aus dem Internet anreichert, um bessere Übersetzungesergebnisse zu erzielen.

Dieser Tage erleben wir, wie tausende von Flüchtlingen zu uns nach Deutschland kommen. Unter den vielen notwendigen Integrationsmaßnahmen wird das Erlernen der deutschen Sprache oft an einer der ersten Stellen genannt.

Wir sollten uns anlässlich des europäischen Tages der Sprachen am 26. September daran erinnern, dass wir auch auf europäischer Ebene einen wichtigen Schritt hin zu erfolgreicher Integration noch nicht angegangen sind. Noch immer sind Sprachbarrieren die stabilsten Grenzen in Europa, die grenzüberschreitenden Handel, Kommunikation und eine echte Integration massiv behindern.

Das Language Technology Lab des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI GmbH) hat dieses Problemfeld bereits seit 2012 in der Weißbuchreihe „Europas Sprachen im digitalen Zeitalter" für bisher 30 europäische Sprachen untersucht. Die Ergebnisse zeigen, dass es insbesondere um die digitale Unterstützung der meisten europäischen Sprachen sehr schlecht steht. Diese Erkenntnis hat die Europäische Union zu dem entscheidenden Schritt bewogen, mit der Connecting Europe Facility (CEF.AT) Plattform künftig verschiedene Anwendungsbereiche im öffentlichen Dienst ohne Sprachbarrieren nutzbar zu machen.

Wie können wir in Zukunft hohe automatische Übersetzungsqualität in beliebigen Anwendungsgebieten sicherstellen?

Das Lernen von Sprachen ist keine leichte Aufgabe. Während es Kindern spielerisch gelingt, ihre eigene oder sogar eine fremde Sprache zu erwerben, indem sie den Erwachsenen zuhören und sich ausprobieren, müssen Erwachsene sich für jede neue Sprache mühsam Grammatikregeln und Vokabeln aneignen.

Computerprogramme wie zum Beispiel die bekannten Online-Übersetzungssysteme lernen Sprache, indem sie große Mengen von mehrsprachigen Texten aus dem Internet mit maschinellen Lernverfahren verarbeiten („Big Data") und so statistische Wahrscheinlichkeiten

lernen, wie bestimmte Wörter und Folgen von Wörter übersetzt werden. Die Systeme können dann ungefähre Übersetzungen erzeugen, deren Qualität je nach Eingabe sehr stark schwankt.

Sprachen sind aber nicht bloß Folgen von Wörtern, sondern haben eine grammatische Struktur, die beispielsweise aus Subjekt, Prädikat und Objekt bestehen kann. Tatsächlich gibt es auch Computersysteme zur maschinellen Übersetzung, die diese Art von Information nutzen. Diese linguistisch motivierten Systeme sind allerdings recht unflexibel und finden heute nur in bestimmten Nischen eine Anwendung. So wird zum Beispiel die Zeitung *La Vanguardia* jede Nacht mit solch einem System von Spanisch ins Katalanische übersetzt und nur noch minimal von menschlichen Übersetzern nachbearbeitet, bevor sie in den Druck geht.

Im Internet gibt es auch semantisches Wissen, etwa in maschinenlesbaren Varianten von Wikipedia, dem sogenannten „Semantic Web“. Diese Ressourcen kodieren Informationen wie „Paris ist die Hauptstadt von Frankreich”, aber auch „Paris ist eine Stadt in Texas”.

Im Rahmen des europäischen Verbundprojektes QTLeap (http://qtleap.eu) entwickelt das Language Technology Lab des DFKI zur Zeit eine hybride Technologie zur maschinellen Übersetzung, die statistische Systeme mit linguistischem Wissen und semantischem Wissen aus dem Internet anreichert, um bessere Übersetzungesergebnisse zu erzielen. Die konkrete Anwendungsdomäne in diesem Projekt ist eine Chat-Hotline, die Nutzeranfragen zu PC-Problemen automatisch in verschiedenen Sprachen beantworten kann.

Die ersten Prototypen sind vielversprechend, aber „wir stehen mit dieser hybriden Forschungsrichtung noch relativ am Anfang“, sagt Professor Hans Uszkoreit, Wissenschaftlicher Direktor und Leiter des Language Technology Labs am DFKI Berlin. „Noch schwieriger als die maschinelle Übersetzung selber ist die Bewertung der Übersetzungsqualität und insbesondere die Diagnose der Schwächen der einzelnen Übersetzungstechnologien. Leider ist es derzeit fast gar nicht möglich, Übersetzungsqualität verlässlich automatisch zu messen. Das ist eine große Herausforderung für das gesamte Forschungsgebiet, an der wir hier am DFKI ebenfalls intensiv arbeiten.“

Während uns unsere Kinder heute fragen, wie wir denn vor zwanzig Jahren ohne Wikipedia unsere Hausaufgaben gemacht haben, werden unsere Enkel vielleicht im Jahr 2035 fragen, wie wir denn damals im Jahr 2015 mit all den Leuten kommuniziert haben, die nach Deutschland gekommen sind, um hier zusammen mit uns ein besseres Leben zu haben – vor der Erfindung des „Omnitranslators”.

Kontakt
Dr. Aljoscha Burchardt
aljoscha.burchardt@dfki.de
tel. 030 23895 1838