56. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 10. bis 12. März 2020

DEUTSCH IN EUROPA

Sprachpolitisch, grammatisch, methodisch

Mittwoch, 11. März 2020, 10:00 Uhr

SPRACHDATEN UND AUTOMATISCHE ÜBERSETZUNG IN EUROPA

Josef van Genabith (Universität des Saarlandes & DFKI)

Sprachtechnologie hat in den letzten 10 Jahren enorme Fortschritte gemacht. In einigen Sprachen, Domänen und Anwendungsbereichen wie Maschine Übersetzung, ASR und Textgenerierung haben die Technologien “human parity“ erreicht. Diese Leistungen beruhen zum größten Teil auf Fortschritten im maschinellen Lernen, im Besonderen im Tiefen Lernen mit neuronalen Netzwerken. Maschinelles Lernen braucht Daten. Um wirklich gute Ergebnisse erzielen zu können sind Datenvolumen, Qualität der Daten sowie Relevanz der Daten (für eine Domäne und Anwendung) entscheidend. Eines der größten Hürden in der Entwicklung von Sprachtechnologien sind Datenengpässe oder, im schlimmsten Fall, Nichtverfügbarkeit von Daten. In diesem Vortrag werde ich exemplarisch zwei Aktivitäten vorstellen, die das Datenproblem in unterschiedlicher Weise angehen. Die Beispiele kommen aus dem Bereich der maschinellen Übersetzung:

- Die EU hat derzeit 28 Mitgliedsländer und 24 offizielle Sprachen. Die EU unterstützt Sprachenvielfalt als Kernteil europäischer Kultur und Identität, möchte aber auch, dass Sprachen nicht zu Barrieren und Silos führen, die die Mobilität von Menschen, den freien Fluss von Ideen, Wirtschaft, Kultur und Administration behindern: “United in Diversity“, „In Vielfalt geeint“. Aus diesem Grund setzt die EU stark auf maschinelle Übersetzung und ist dabei, diese nicht nur für EU interne Bedürfnisse, sondern auch für öffentliche Verwaltungen sowie NGOs in den EU Mitgliedsländern zur Verfügung zu stellen. Dazu fehlen der EU aber die geeigneten Trainingsdaten. Aus diesem Grund haben wir seit 5 Jahren im Rahmen des CEF (Connecting Europe Facilty) Programms ein europäisches Netzwerk (ELRC European Language Resource Coordination http://www.lr-coordination.eu/) aufgebaut, das in ganz Europa sowie in den CEF affillierten Ländern (Norwegen und Island) Strukturen aufgebaut hat, Übersetzungsdaten von öffentlichen Institutionen und NGOs zu sammeln: „Supporting your language supports Europe, and supporting Europe is supporting your language.“

- Für die große Mehrheit der 7000+ lebenden Sprachen der Welt, und insbesondere auch für viele sogenannte kleinere Sprachen Europas, gibt es nicht genug oder meistens sogar keine parallelen Trainingsdaten für maschinelle Übersetzung. In der Forschung gibt es viele Ansätze, die versuchen aus dieser Sachlage das Beste zu machen: Übersetzung durch eine Pivotsprache, unüberwachte (unsupervised) maschinelle Übersetzung, massiv multilinguale (massively multi-lingual) maschinelle Übersetzung, Transfer-Lernen (transfer-learning) basierte maschinelle Übersetzung, usw. Hier werde ich kurz selbst-überwachte (self-supervised) maschinelle Übersetzung [Ruiters et al. 2019] vorstellen, das gleichzeitig Übersetzten und Extraktion von parallelen Daten aus vergleichbaren Daten (comparable data, e.g. wikipedia) lernt.

Literatur:

[Ruiters et al. 2019] Dana Ruiter, Cristina España-Bonet and Josef van Genabith. Self-Supervised Neural Machine Translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Pages 1828-1834. Florence. Italy. August 2019.