satw it's all about technology

Die automatische Verarbeitung von natürlicher Sprache

Forschungsresultate des Labors für verteilte Informationssysteme (LSIR) der EPF Lausanne

Die automatische Verarbeitung von natürlicher Sprache (nachfolgend LDV für linguistische Datenverarbeitung) ist eine Forschungsrichtung, welche versucht, Computern die menschliche bzw. «natürliche» Sprache beizubringen. Dies kann schriftliche oder mündliche Sprache sein. Durch LDV wird erreicht, dass eine Maschine die Anfrage eines Menschen versteht, diese interpretiert und eine adäquate Antwort gibt.

LDV ist ein multidisziplinäres Forschungsgebiet, welches die Kenntnisse der Informatik mit denen der Linguistik und der künstlichen Intelligenz verbindet. Dank der wachsenden Rechenleistung von Computern, der grossen Verfügbarkeit von Open-Source-Daten und der Perfektionierung der Algorithmen zum maschinellen Lernen (*machine learning*) und zum Deep Learning hat das Forschungsgebiet von LDV in den letzten Jahren ein beispielloses Wachstum erlebt. LDV ist bereits in unterschiedlichen Anwendungen des täglichen Lebens präsent: Suchmaschinen, Sprachassistenten, virtuelle Assistenten (*Chatbots*) oder Spamfilter sind der breiten Bevölkerung bestens bekannt.

Das Labor für verteilte Informationssysteme der EPF Lausanne untersucht derzeit, wie LDV an spezialisierte Gebiete und an andere Sprachen als Englisch angepasst werden kann. Die Aufgabe der Gruppe besteht darin, aus der enormen Menge der im Internet verfügbaren Daten zuverlässige Informationen zu erstellen – eine grosse Herausforderung in der heutigen Informationsgesellschaft. Die Forschenden entwickeln Methoden und Systeme zur Umwandlung unstrukturierter, heterogener und unzuverlässiger Daten in aussagekräftige, zuverlässige und verständliche Informationen.

Von rekurrenten neuronalen Netzen zu *Transformern*

Damit eine Maschine eine Sprache lernt, muss natürliche Sprache in mathematische Formeln überführt werden. Der Ansatz ist probabilistisch und basiert auf Algorithmen. Die ersten Ansätze dazu beruhten auf der Art, wie wir lesen: eine sequenzielle Verarbeitung der aufgenommenen Information, ein Wort nach dem anderen. Dieses sogenannte «rekurrente» System basiert auf den rekurrenten neuronalen Netzwerken (kurz RNN für *Recurrent Neutral Networks*). Dabei wird das Lernen in Form eines Zustandsvektors konstruiert, der eine Repräsentation der bisher im Satz angetroffenen Daten enthält. Bei jedem neuen Wort wird der Zustandsvektor angepasst, um den neuen Zustand darzustellen. Dieser Ansatz erlaubt ein Basisverständnis und die Ausführung von simplen Aufgaben. Die Abbildung 1 zeigt eine grafische Darstellung eines RNN.

Herkömmliche RNN sind anfällig für Gradientenverluste, was zu einer Verschlechterung des Informationsgehalts führt, wenn die RNN an Grösse und Komplexität zunehmen. In den letzten zehn Jahren wurde es durch die Fortschritte im Deep Learning und durch

das Kapazitätswachstum möglich, sogenannte *Transformer* Modelle zu entwickeln. Diese lösten eine regelrechte Revolution aus. Anders als die RNN verarbeiten *Transformer* die Informationen nicht in der Reihenfolge in der sie erscheinen, vom Satzanfang bis zum Ende. Stattdessen identifizieren sie den Kontext, welcher jedem Wort im Satz eine Bedeutung verleiht. Dies geschieht durch einen Aufmerksamkeits-Mechanismus, welcher die globalen Zusammenhänge herstellt. Dieses Modell hat die Arbeitsweise mit textuellen Daten fundamental verändert und es ermöglicht, ein Training der Algorithmen mit grösseren Datensätzen als früher durchzuführen. All dies hat zur Entwicklung von vortrainierten Systemen wie BERT geführt, ein Akronym für *Bidirectional Encoder Representations from Transformers,* welches von Google entwickelt und 2018 veröffentlicht wurde. BERT ist öffentlich zugänglich und kann von den Nutzerinnen und Nutzer verwendet und angepasst werden.

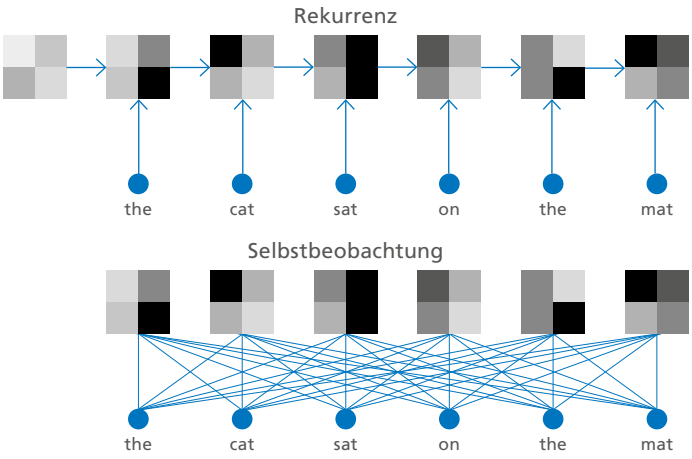

Abbildung 1
Vergleich eines rekurrenten neuronalen Netzwerks (RNN) mit einem Aufmerksamkeits-Mechanismus Modell. Beim RNN steht jedes Wort in Abhängigkeit zu den vorherigen Wörtern. Dies macht es für das Modell schwierig, Zusammenhänge in längeren Sätzen zu erkennen. Beim Aufmerksamkeits-Mechanismus Modell ist jedes Wort mit allen weiteren Wörtern der gesamten Sequenz verbunden. So können auch Zusammenhänge zwischen weit entfernten Wörtern modelliert werden.

Daten und Lernprozess

Um LDV-Algorithmen zu trainieren, wird eine enorme Datenmenge benötigt. Das Internet, insbesondere Wikipedia, und Online-Bibliotheken sind hier sehr ergiebige Quellen für grosse Mengen an allgemeinen Texten.

Diese Texte dürfen nicht einfach so, wie sie sind, verarbeitet werden. Sie müssen durch Verarbeitungstechniken wie zum Beispiel die *Tokenisierung*, sprich die Segmentierung der Sprache in kleine lexikalische Einheiten (*Token*), in verwertbare Daten umgewandelt werden.

Der Lernprozess besteht dann daraus, den Algorithmus mit diesen Daten zu füttern und ihm spezifische Aufgaben zu stellen, die es ihm ermöglichen, sein Begriffsvermögen anzupassen. BERT wurde mit Aufgaben zur maskierten Sprachmodellierung (*Masked Language Modeling*) und zur Vorhersage des nächsten Satzes vortrainiert. Der Ausbildungsprozess besteht dabei aus zwei Phasen. Beim ersten Training wird das Modell mit unmarkierten Daten trainiert. Für die Feinabstimmung werden dann alle Parameter anhand von markierten Daten angepasst.

Fallstudie I:
Anpassung des BERT-Modells an die Rechtssprache

BERT wurde mit einem allgemeinen Sprachkorpus trainiert. In vielen Fällen ist es möglich, BERT unverändert für linguistische Aufgaben zu verwenden. Studien haben jedoch gezeigt, dass die Leistungen eines vortrainierten Modells verbessert werden, wenn es an den spezifischen Bereich seiner Zielaufgabe angepasst wird.

Abbildung 2 zeigt ein Beispiel für die Verteilung und Überschneidung von Daten aus einem spezialisierten Bereich mit den Daten eines allgemeinen Bereichs sowie den aufgabenspezifischen Daten (in Grau). Überschneidungen sind möglich, aber nicht notwendig, was die Bedeutung eines domänenspezifischen Trainings verdeutlicht.

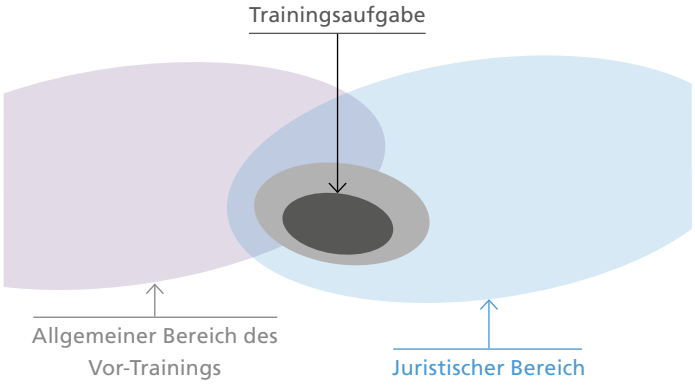

Abbildung 2
Illustration der Datenverteilung. Die Daten zu der Aufgabe verteilen sich auf erkennbare Art, und sie werden in der Regel nicht zufällig einer grösseren Verteilung (hellgraue Ellipse) innerhalb eines noch grösseren Zielbereichs entnommen. Dieser deckt sich nicht zwingend mit den Bereichen, die im anfänglichen Vor-Trainings-Bereich enthalten sind, obwohl Überschneidungen möglich sind (nach Gururangan et al., ACL 2020).

Es gibt eine Reihe von wichtigen Merkmalen bei der Verarbeitung von juristischer Sprache, die eine Herausforderung darstellt. Die Rechtssprache ist im Vergleich zur «normalen» Sprache komplex und formal. Die Menge der öffentlich zugänglichen Textdaten ist hier begrenzt, da viele Dokumente vertraulich sind. Schliesslich kann die Verwendung von LDV im rechtlichen, regulatorischen oder politischen Bereich schwerwiegende Folgen haben, was es nicht erlaubt, Fehler zu machen.

Das Ziel der vom LSIR an der EPFL durchgeführten Studie war es, zu bewerten, inwieweit das adaptive Vortraining eines von BERT abgeleiteten Modells im juristischen Bereich dessen Fähigkeit verbessert, eine bestimmte Klassifizierungsaufgabe zu erfüllen.

In einem ersten Schritt sammelten die Forschenden mehrere zehntausend frei zugängliche Dokumente des öffentlichen europäischen Rechts in Englisch, Französisch, Deutsch und Italienisch. Die Grösse des Textkorpus in jeder Sprache war mit 5 bis 6 GB relativ gering im Vergleich zu den Korpora, die in anderen Bereichen verwendet werden (Tabelle 1).

Bereich	Name des Korpus	Grösse (in GB)	Grösse (in GB)
Biomedizin	S2ORC	7,55	47
Informatik	S2ORC	8,10	48
Nachrichten/News	REALNEWS	6,66	39
Meinungen von KonsumentInnen	AMAZON REVIEWS	2,11	11
Jurisprudenz	LEGAL-ROBERTA	1,01	4,9

Tabelle 1
Merkmale von verschiedenen Datenkorpora aus unterschiedlichen Bereichen. Diese unmarkierten Daten ermöglichen ein unüberwachtes Lernen zur Anpassung von Modellen an spezifische Bereiche.

Diese Daten wurden für das adaptive Vortraining des RoBERTa-Modells verwendet – einer Version von BERT, die auf einem zehnmal grösseren Datenkorpus vortrainiert wurde. Dazu wurde eine maskierte Sprachmodellierungsaufgabe angewendet.

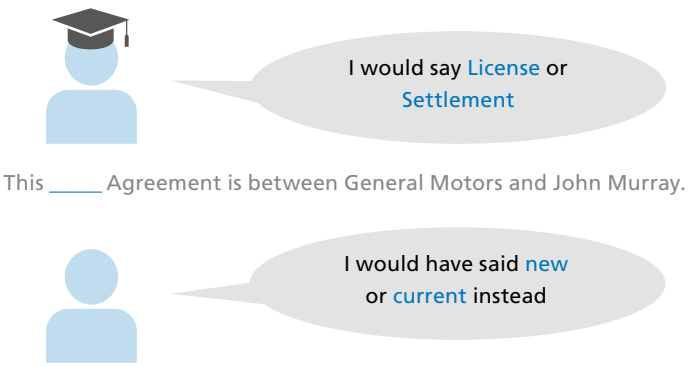

Abbildung 3
Beispiel für den Unterschied zwischen der Erkennung von ausgeblendeten Wörtern beim LegalRoBERTa Modell (oben), das für den juristischen Bereich adaptiert wurde, und dem BERT Modell (unten), welches mit einem Korpus der allgemeinen Sprache trainiert wurde.

Das Lernen im juristischen Bereich wurde durch überwachtes Lernen verfeinert, d.h. Lernen mittels Daten, die in einer aufgabenadaptiven Weise annotiert wurden. Darin enthalten waren Aufgaben zur Dokumentenklassifikation und zum Suchen von juristischen Phrasen in einem Datensatz von fast 4'000 kommentierten Rechtsfällen.

Das Modell wurde dann anhand einer Auswahl von 600 Dokumenten bewertet, die zuvor von Expertinnen und Experten in zwei Kategorien eingeteilt wurden: rechtlich bindend und/oder rechtlich nicht bindend.

Die Ergebnisse der Textklassifizierungsaufgabe wurden mit denen von RoBERTa und einer anderen, auf den juristischen Bereich vortrainierten Version, BERT, verglichen. Ein typisches Ergebnis derselben Aufgabe, die diesen drei Modellen zugewiesen wurde, ist in Tabelle 2 dargestellt. Tabelle 2 zeigt die Ergebnisse der Klassifizierungsaufgabe gemäss den beiden Metriken, die üblicherweise zur Bewertung von Klassifizierungsaufgaben verwendet werden: Präzision P (oder positiver prädiktiver Wert) und Recall R (oder Sensitivität).

Modell	Recall@10	Mittlerer Median-Rang
BERT	43,7%	14
RoBERTa	46,6%	13
LegalRoBERTa	**48,3%**	**12**

Tabelle 2
Resultate der Aufgabe der Informationssuche. Ausgehend von einem juristischen Ausdruck, der ein langes Dokument zusammenfasst, misst der Recall@10 die Wahrscheinlichkeit ein relevantes Dokument in den ersten 10 Resultaten aufzufinden.

LegalRoBERTa übertraf die anderen beiden Modelle bei den meisten Metriken, trotz der geringen Grösse des Dokumentenkorpus, der für das domänenspezifische Training verwendet wurde.

Fallstudie II: mehrsprachige Anpassung

Die meisten LDV-Modelle sind mit englischsprachigen Korpora vortrainiert. Es gibt eine mehrsprachige Version von BERT, M-BERT, die auf 104 verfügbaren Sprachen von Wikipedia trainiert ist, aber diese Version ist nicht an die Rechtssprache angepasst.

Um M-BERT für den juristischen Bereich in drei Schweizer Landessprachen zu trainieren, gibt es zwei Möglichkeiten: die Übersetzung der Dokumente auf Englisch mit Hilfe von Standardübersetzern, zum Beispiel Google Translate, oder die Verwendung von Sprachadaptern.

Der Sprachadapter wird durch unüberwachtes Lernen mit einer Reihe von Dokumenten in Englisch, Französisch, Deutsch und Italienisch trainiert. Dann wird ein Aufgabenadapter mit einer Zero-Shot-Lernmethode trainiert, die darin besteht, das Modell anzuweisen, Objekte ohne vorherige Informationen oder mit Hilfe von Zwischenrepräsentationen zu klassifizieren. In diesem Fall bestand die Aufgabe darin, französische und deutsche Dokumente auf der Grundlage von englischen Dokumenten zu klassifizieren.

In seiner Fallstudie hat das Labor für verteilte Informationssysteme der EPFL diese beiden Ansätze für französische und deutsche Dokumente verglichen. Die Ergebnisse des Experiments zeigten, dass die Verwendung von Sprachadaptern bessere Ergebnisse liefern kann als die Verwendung übersetzter Daten.

Die SATW fördert Projekte im Bereich Künstliche Intelligenz

Viele Organisationen beschäftigen sich mit der Frage, wie sie mit Hilfe automatisierter Textanalyse ihre umfassenden Dokument-Ressourcen besser nutzen können. Tech-Giganten treiben die Entwicklung durch den Einsatz künstlicher Intelligenz voran. Die Rahmenbedingungen variieren aber stark. So verfügen diese Tech-Giganten über riesige Datenbestände und umfangreiches Feedback durch die Nutzenden, die zum Training von Textanalyse-Algorithmen eingesetzt werden können.

In vielen industriellen und administrativen Anwendungen – so auch in der öffentlichen Verwaltung – sind die Datenbestände jedoch beschränkt, und das Feedback kommt höchstens von Expertinnen und Experten. Für diese Organisationen stellt sich daher die Frage, inwiefern der globale technische Fortschritt in der Dokument-Analyse auch in spezialisierten Anwendungsfeldern gewinnbringend eingesetzt werden kann, um die verfügbaren Informationen besser nutzbar zu machen und die Endnutzerinnen und Endnutzer von Routineaufgaben zu entlasten.

Die SATW unterstützte ein Team von EPFL-Forschenden unter der Leitung des SATW-Experten Prof. Karl Aberer in der Planung, Umsetzung und Bekanntmachung des Pilotprojekts, das in diesem Factsheet vorgestellt wird. Der Use Case wurde im Auftrag des Eidgenössischen Departments für auswärtige Angelegenheiten (EDA) durchgeführt und steht repräsentativ für vergleichbare Situationen in Industrie und Verwaltung. Besonderes Interesse bestand an der automatisierten Textanalyse von internationalen Vereinbarungen und legalen Verpflichtungen, um diese leichter zu überwachen und eine qualitative Verbesserung des Monitorings herbeizuführen.

Für den Bereich Künstliche Intelligenz (KI) verfügt die SATW über eine Themenplattform und ein Schwerpunktprogramm, organisiert regelmässig Dialogformate und gibt Publikationen heraus – beides in Zusammenarbeit mit Expertinnen und Experten wie Partnerorganisationen. Die SATW ist auch im Bereich Daten sehr aktiv, denn ohne Daten funktioniert KI nicht. Im Frühling 2021 hat die SATW gemeinsam mit der Direktion für Völkerrecht des EDA, dem Bundesamt für Kommunikation und der Swiss Data Alliance das nationale «Netzwerk Digitale Selbstbestimmung» gegründet, um Bürgerinnen und Bürgern, Unternehmen und öffentlichen Einrichtungen eine Nutzung der Datenwirtschaft auf der Basis freiheitlich-demokratischer Grundwerte zu ermöglichen.

Impressum

AutorInnen: Anne May (Radar RP) und Rémi Lebret (EPFL) | **Projektleiter:** Karl Aberer (EPFL) | **Projektteam:** Rémi Lebret (EPFL), Saibo Geng (EPFL), Xiangcheng Cao (EPFL), Stephan Michel (EDA), Claude Schenker (EDA), Manuel Kugler (SATW), Adriana Cantaluppi (SATW) | **Übersetzung:** Life Science Communication AG | **Redaktion:** Beatrice Huber, Esther Lombardini und Alexandre Luyet

Vollständige Publikation zum Projekt

– Geng, Saibo et al. «Legal Transformer Models May Not Always Help». arXiv preprint arXiv:2109.06862 (2021).

Weitere Publikationen zum Thema

– Chalkidis, Ilias, et al. «LEGAL-BERT: The muppets straight out of law school». arXiv preprint arXiv:2010.02559 (2020).
– Devlin, Jacob, et al. «Bert: Pre-training of deep bidirectional transformers for language understanding». arXiv preprint arXiv:1810.04805 (2018).
– Gururangan, Suchin, et al. «Don't stop pretraining: adapt language models to domains and tasks». arXiv preprint arXiv:2004.10964 (2020).
– Liu, Yinhan, et al. «Roberta: A robustly optimized bert pretraining approach». arXiv preprint arXiv:1907.11692 (2019).
– Pfeiffer, Jonas, et al. «Mad-x: An adapter-based framework for multi-task cross-lingual transfer». arXiv preprint arXiv:2005.00052 (2020).

satw it's all about technology

Schweizerische Akademie der Technischen Wissenschaften SATW
St. Annagasse 18 | 8001 Zürich | 044 226 50 11 | info@satw.ch | www.satw.ch

Bild: Adobe Stock