Künstliche Intelligenz

Stefan Geißler*

Natürliche Sprachverarbeitung und Künstliche Intelligenz – ein wachsender Markt mit vielen Chancen

Das Beispiel Kairntech

<https://doi.org/10.1515/iwp-2020-2079>

Vor rund einem Jahr haben wir an dieser Stelle¹ die aufregende Dynamik auf den Gebieten der Natürlichen Sprachverarbeitung (NLP) und der Künstlichen Intelligenz (KI) beschrieben: Seit einigen Jahren sorgen Fortschritte in den algorithmischen Grundlagen, in der einsetzbaren Rechenleistung sowie in der Verfügbarkeit von großen Datenmengen für immer leistungsfähigere Systeme. NLP-Anwendungen seien damit mehr denn je reif für den praktischen Einsatz, hatten wir argumentiert. Diese Entwicklung verfolgen wir bei Kairntech nicht allein als interessierte Beobachter, sondern sie stellt die Grundlage unserer Arbeit dar, mit der wir NLP- und KI-Ansätze zur Anwendung auf konkreten geschäftskritischen Prozessen entwickeln und einsetzen.

Experten gehen auch für die kommenden Jahre von einem anhaltenden Wachstum des weltweiten Marktes für NLP aus: Mit einem durchschnittlichen Wachstum von über 20 Prozent pro Jahr werde der Markt bis 2025 auf geschätzte 6,24 Milliarden US-\$ anwachsen². Im Bereich der Forschung ist das Wachstum sogar noch stürmischer: So ist die Zahl der Einreichungen zur ACL-Konferenz, dem vielleicht wichtigsten jährlichen Event in diesem Gebiet, von 2018 bis 2019 um ganze 75 Prozent angestiegen³.

Im vorliegenden Text wollen wir die Richtung, die wir bei Kairntech mit der Gründung vor einem Jahr eingeschlagen haben, beschreiben sowie von ersten Erfolgen auf diesem Weg berichten.

KI und die Kultur der Wissensteilung

Künstliche Intelligenz gilt weltweit als einer der ausschlaggebenden Wachstumsbereiche, die in Zukunft die Wettbewerbsfähigkeit von Volkswirtschaften entschieden werden. Die Gewichtsverschiebung im Bereich Automobil kann als Beispiel herangezogen werden: Hier nimmt der Anteil der Wertschöpfung durch die Motorenentwicklung, bei der Deutschland traditionell eine sehr starke Stellung hat, mit dem Aufkommen der Elektromobilität ab, während die Software, etwa bei Vernetzung und autonomem Fahren, stark an Bedeutung gewinnt.

Ein bemerkenswerter Aspekt der KI ist die ausgeprägte Tendenz der Akteure zum Teilen („Sharing“), die maßgeblich mitverantwortlich für die raschen Fortschritte ist⁴. Konzerne wie akademische Akteure teilen intensiv relevante Software-Bausteine. Führende Forscher wie Andrew Ng und Sebastian Thrun fördern Initiativen wie Coursera und Udacity⁵, die weltweit den Zugang zu relevanten Lerninhalten in KI – und vielen anderen Wissensgebieten ermöglichen. Und Plattformen wie arxiv und github⁶ versetzen wissenshungrige Menschen in aller Welt in die Lage, rasch auf Tuchfühlung mit dem Stand der Forschung zu gehen. Wie bemerkenswert diese Kultur des Teilens ist, wird deutlich, wenn man sie mit anderen Technologiefeldern vergleicht: Schwer vorstellbar, dass eine Firma oder ein Forschungsteam einen wichtigen Fortschritt hin zu noch effizienteren Batterien oder schädlingsresistentem

1 Stefan Geißler, Information – Wissenschaft & Praxis 70, 2–3, 134–140, doi: <https://doi.org/10.1515/iwp-2019-2007>.

2 Vgl. Kenneth Research “Natural Language Processing (NLP) Market Analysis, Size, Share, Growth, Trends and Forecast to 2025”. America News Hour, 2020.

3 <https://www.mihaileric.com/posts/nlp-trends-acl-2019/> [20.1.2020].

*Kontaktperson: Stefan Geißler, Kairntech Deutschland UG, Schützenhausstr 28, 69151 Neckargemünd, E-Mail: stefan.geissler@kairntech.com, www.kairntech.com

4 Ein systematischer Vergleich der „Sharing“-Kultur in unterschiedlichen Fachgebieten ist uns nicht bekannt. Die weit fortgeschrittene Tendenz zur Vereinheitlichung von Evaluationsdaten sowie der IT-Umgebung vieler NLP/KI-Arbeiten – Linux, Python, Tensorflow, u. ä. – stützt zumindest die Annahme, dass diese Kultur in NLP/KI in besonderem Maße ausgeprägt ist, selbst wenn natürlich bei weitem nicht bei allen Veröffentlichungen der zugrundeliegende Code und die Daten umgehend frei verfügbar gemacht werden.

5 <https://de.coursera.org/>, <https://www.udacity.com/> [20.1.2020].

6 <https://www.udacity.com/>, <https://github.com/> [20.1.2020].

Saatgut umgehend samt den nötigen Hinweisen zum Nachbauen in der Public Domain veröffentlicht.

Während sich aus dieser Sicht die Hürden für einen Einstieg von neuen Akteuren im Bereich KI abgesenkt haben, gibt es andererseits eine Verlagerung der wettbewerbsrelevanten Faktoren: Von den Algorithmen hin zu der Frage nach der Verfügbarkeit von Daten, dem nötigen Domänenwissen, um die Verfahren sinnvoll für geschäftsrelevante Fragen einsetzen zu können, sowie zu der Frage der Integration von KI in umfassendere IT-Umgebungen.

Diese Überlegungen haben von Beginn an die Ausrichtung von Kairntech und die unserer Softwareentwicklung und unserer Plattform *Sherpa* bestimmt.

Datenverfügbarkeit und Transfer-Learning

Denn wenn die grundlegenden Algorithmen frei verfügbar sind, jedoch erst mit geeigneten Trainingsdaten ihr volles Potenzial entfalten können, bedeutet dies, dass Verfügbarkeit, Qualität, Umfang von Trainingsdaten sowie die Kosten, zu denen diese erstellt werden können, oft die entscheidenden Faktoren sind, die über den Erfolg oder Misserfolg des Einsatzes von KI im Unternehmen entscheiden. Ein zentrales Anwendungsfeld des Instruments *Sherpa* ist daher die effiziente Erstellung von Trainingsdatensätzen. Die Einsetzbarkeit von Umgebungen zur Annotierung von Daten für Dokumentenanalyse-Projekte spielt dabei einem Trend in die Karten, der in der Machine Learning (ML)-Community in den letzten Jahren stark an Bedeutung zugenommen hat: Das Transfer-Learning⁷. Während das Training von Deep Learning Modellen normalerweise sehr große Mengen von annotierten Daten voraussetzt, zeigt sich inzwischen, dass es oft erfolversprechend ist, komplexe Modelle, die mit großem Aufwand vortrainiert wurden (pretrained), anschließend an konkrete Fragestellungen zu adaptieren. Dabei werden sie auf vergleichsweise überschaubaren Datenmengen für diese Aufgabe angepasst – die vorab gelernten Fähigkeiten können quasi auf die jeweils anstehende Anforderung transferiert werden. Modelle wie *BERT*⁸, die mit für die meisten Teams unerreichbaren Aufwänden trainiert wurden, können damit zur Grundlage eigener Arbeiten werden. Entsprechend gute Erfahrungen wurden auch bei Kairntech mit dem Einsatz von *BERT* und konkreten Ausprägungen davon wie *SciBERT*⁹ (angepasst auf die Analyse wissenschaftlicher Inhalte) gemacht.

Sharing

Wie in wohl wenig anderen Technologiefeldern ist im Bereich KI/NLP eine Kultur des Teilens entstanden, die es neuen Mitspielern vergleichsweise einfach macht, rasch auf Arbeiten von Dritten aufzubauen: Programmbibliotheken wie *Tensorflow* oder vorberechnete Modelle wie *BERT* sind in der Public Domain, Plattformen wie *github* fördern die Verbreitung, oft gleich zusammen mit der Dokumentation, den Originalpublikationen sowie häufig sogar Tutorial-Material. Kairntech ist nicht nur intensiver Nutznießer dieses Ökosystems, sondern trägt wiederum mit eigenen Arbeiten dazu bei: Komponenten wie *DELFT* (Framework zu Erstellung von Deep Learning Modellen¹⁰) oder *Grobid*¹¹ (Machine-Learning-getriebene Strukturierung von Dokumenten) stammen von Mitgliedern des Kairntech-Teams und sind nicht nur ebenfalls in der Public Domain verfügbar, sondern bereits im produktiven Einsatz in großen laufenden Projekten.

Demokratisierung

Mit dem Begriff Demokratisierung verbindet sich der Anspruch, relevante Fortschritte nicht länger nur einem kleinen Kreis von Expertinnen und Experten verfügbar zu machen. Vielmehr sollen durch die Einbettung in einfache IT-Umgebungen und Benutzerinterfaces auch Personen ohne ausgeprägte IT-Spezialkenntnisse, in diesem Falle also nicht nur Datenwissenschaftler und Machine-Learning-Experten, sondern auch Fachleute aus anderen Disziplinen, wie z. B. der Biologie, Jura, Marketing oder Verwaltung, Zugang haben. Bei der Entwicklung des Kairntech *Sherpa* war insbesondere dieser Aspekt prägend. Die Erstellung, Bewertung und Produktiv-Setzung von leistungsfähigen ML-Modellen sollte auch für diese zweite Personengruppe möglich sein.

Der Sherpa

Die Schlussfolgerungen, die wir bei Kairntech aus den oben angeführten Trends gezogen haben, sind in die Entwicklung einer Plattform eingegangen, mit der wir sowohl das Thema Datenverfügbarkeit als auch die breite Einsetzbarkeit von leistungsfähigen maschinellen Lernverfahren für Domänenexperten adressieren. Was die Datenverfüg-

⁷ <https://www.aclweb.org/anthology/N19-5004.pdf> [20.1.2020].

⁸ <https://github.com/google-research/bert> [20.1.2020].

⁹ <https://arxiv.org/abs/1903.10676> [20.1.2020].

¹⁰ <https://github.com/kermitt2/delft> [20.1.2020].

¹¹ <https://github.com/kermitt2/grobid> [20.1.2020].

Abbildung 1: Sherpa ist eine webbasierte Umgebung, die dazu einlädt Text mit Entitäten zu annotieren. Im Hintergrund liefert ein Machine Learning Prozess kontinuierlich immer detaillierte Prognosen bis die gewünschte Qualität erreicht ist.

barkeit angeht, so ist die Beobachtung, dass trotz scheinbar überreichlich vorhandenen Daten zu praktisch allen denkbaren Themen auch in der Public Domain in vielen Fällen zu Beginn eines ML-Projektes zunächst die Verfügbarkeit von geeigneten Trainingsdaten sichergestellt werden muss. Oft müssen dafür signifikante Aufwände bei der manuellen Aufbereitung (Annotierung, Labelling) von Daten getrieben werden. Ob dieser Prozess effizient gestaltet werden kann, ist oft von entscheidender Bedeutung für die Umsetzbarkeit des Projekts: Aufwände von mehreren Wochen alleine zur Erstellung von Trainingsdaten können ein Projekt rasch unrealistisch werden lassen, während eine Umgebung, die durch geeignete Benutzerführung und weitere Maßnahmen (z. B. *Active Learning*¹²) die Aufwände signifikant reduzieren hilft, ein wichtiger Baustein für ein erfolgreiches Projekts ist.

Damit Hand in Hand geht die Ausrichtung der gewählten Umgebung an den Kompetenzen und Fähigkeiten von Domänenexpertinnen und -experten statt an denen von Fachleuten in Data Science: Obwohl nämlich populäre Programmier-Umgebungen wie *Tensorflow* und *PyTorch*¹³ rasche Fortschritte hin zu immer größerer Abstraktheit und Unterstützung der Benutzer machen, bleibt ihr produktiver Einsatz Anwendern mit Kenntnissen in Programmierung und anderen IT-lastigen Themen vorbehalten. Im Unternehmensalltag sind es dagegen oftmals die Fachleute in den jeweiligen Fachdomänen (Medizin, Recht, Biologie, Marketing, ...), die detaillierte Kenntnis hinsichtlich der Daten, jedoch nicht notwendigerweise einen Hintergrund in Data Science, haben. Gerade an diese Nutzergruppen richtet sich der *Sherpa*.

Beispiele aus der Praxis

Zwei Beispiele aus der Praxis vermitteln einen Eindruck davon, wo Lösungen mit den beschriebenen Eigenschaften heute eingesetzt werden: Das Pharmaunternehmen Boehringer Ingelheim beobachtet wie alle Akteure in diesem Bereich, dass die Pharma-Industrie heute in hohem Maße wissensgetrieben ist. Informationen aus einer oft überwältigenden Anzahl von Dokumenten sind auszuwerten und zu beachten und können Einfluss auf den Fortgang oder gar das Wohl und Wehe von milliardenteuren Medikamentenentwicklungsprojekten haben. Eine eigene Abteilung im Unternehmen, das Scientific Information Center (SIC), hat daher die Aufgabe, die Datenflut für die Kolleginnen und Kollegen im Unternehmen in geeigneter Form aufzubereiten. Der Abteilungsleiter des SIC, Dr. Karlheinz Spenny, stellt hierzu fest: „Ein Großteil des verfügbaren Wissens liegt in Form von unstrukturiertem Text vor, in wissenschaftlichen Veröffentlichungen, Patenten, Konferenzbeiträgen oder auf Websites. Nötig ist es oft, implizite Beziehungen z.B. zwischen Genen und zellulären Targets zu ermitteln. Dieses Wissen ist essenziell in der Medikamentenentwicklung und leistungsfähige AI und NLP Lösungen sind entscheidend dafür, diese Art von Informationen effizient und rasch für unsere Wissenschaftler zu verarbeiten.“¹⁴ Als wichtigen Baustein bei dieser Aufgabe hat man sich bei Boehringer Ingelheim daher für den Einsatz des Kairntech *Sherpa* entschieden.

Von einer anderen Art ist das Szenario für den Einsatz von Kairntech beim französischen Informationsanbieter SEALK: Dort werden Informationen aus dem Netz gezielt unter dem Blickwinkel ihrer Relevanz für Fragen rund um Firmenzusammenschlüsse (Mergers & Acquisitions) gesammelt, analysiert und angeboten. Die Fähigkeiten von

¹² Samut, Claude and Geoffrey I. Webb (Hrsg.): *Encyclopedia of Machine Learning*, Springer 2010, doi.org/10.1007/978-0-387-30164-8_6

¹³ <https://www.tensorflow.org/>, <https://pytorch.org/> [20.1.2020].

¹⁴ Vgl. <https://www.kairntech.com/articles/dec2019.html> [20.1.2020].

Kairntech zur Erkennung, Disambiguierung und Verlinkung von benannten Entitäten wie Firmen und Organisationen, Personen, Orten und anderen Typen sind hier wichtige Bausteine des entsprechenden Angebots von sealk.co.

Ausblick

Anwendungen aus dem Bereich NLP haben in den letzten Jahren, nicht zuletzt dank der Fortschritte bei maschinellen Lernverfahren, massiv an Alltagstauglichkeit gewonnen. Themen wie Spracherkennung und maschinelle Übersetzung, aber auch Kategorisierung, die Erkennung von Entitäten oder Sentiment-Analyse können in einer Qualität adressiert werden, die der Bearbeitung durch Menschen immer näherkommt.

Es verbleiben jedoch noch zahlreiche Fragen, bei denen NLP derzeit noch weit davon entfernt ist, industriell einsetzbare Lösungen anbieten zu können, beispielsweise Inferenzen zu prozessieren, also das automatische Ziehen von Schlüssen aus in textueller Form vorliegendem Wissen. Dies wäre eine überaus wichtige Funktion, ist jedoch derzeit längst noch nicht befriedigend gelöst. Der Hoffnung, der in den letzten Jahren um sich greifenden Flut von Fake News ließe sich mit automatischen Systemen Einhalt gebieten, muss ebenfalls ein Dämpfer verpasst werden. Und die derzeit kaum in Ansätzen mögliche Nachvollziehbarkeit der Entscheidungen von komplexen Neuronalen Netzen (NN), die oftmals praktisch als Black Box gesehen werden müssen, ist eine weitere Hürde hin zu noch größerer Verbreitung von NN-basierter NLP. Eine neue Forschungsrichtung namens X-AI („Explainable AI“) ist daran, diesen Mangel zu beheben.

Generell mehren sich trotz der unbestreitbaren Erfolge von NLP und anderen KI-Themen, die auf dem Paradigma Deep Learning fußen, die Stimmen, die in Frage stellen, ob der Fortschritt ähnlich ungebremsst auch die noch ausstehenden Aufgaben anzugehen erlauben wird, oder ob Deep Learning als Werkzeug nicht absehbar an seine Grenzen geraten wird oder bereits geraten ist und komplettiert werden muss durch andere Methoden.¹⁵

Fazit

Anwendungen im Bereich NLP erleben derzeit ähnlich tiefgreifende Fortschritte wie AI Anwendungen zu ande-

ren Themen wie Bildverarbeitung, Robotik oder Agenten in Spielsituationen. Eine wichtige Zutat hierbei sind moderne Lernverfahren, die bei vielen Teilaufgaben inzwischen die Standards hinsichtlich Qualität setzen. Während es dabei weiter einer kleinen Schar von Spezialisten vorbehalten bleiben wird, diese Fortschritte im Detail nachvollziehen oder gar dazu beitragen zu können, erlauben leistungsfähige Programmierbibliotheken und Entwicklungswerkzeuge einer größer werdenden Anzahl von Menschen, diese Methoden zu nutzen und für eigene Anforderungen einzusetzen.

Quellen

- Marcus, Gary und Ernest Davis: „Rebooting AI“, Pantheon, 2019, ISBN 9781524748258.
- Stefan Geißler: „Maschinelles Lernen und NLP: Reif für die industrielle Anwendung!“, Information – Wissenschaft & Praxis 70, 2–3, 134–140, doi: <https://doi.org/10.1515/iwp-2019-2007>.
- Kenneth Research “Natural Language Processing (NLP) Market Analysis, Size, Share, Growth, Trends and Forecast to 2025”. America News Hour, 2020.
- <https://www.kairntech.com/articles/dec2019.html> [20.1.2020].
- <https://www.mihaileric.com/posts/nlp-trends-acl-2019/> [20.1.2020].
- <https://de.coursera.org/>, <https://www.udacity.com/> [20.1.2020].
- <https://www.udacity.com/>, <https://github.com/> [20.1.2020].
- <https://www.aclweb.org/anthology/N19-5004.pdf> [20.1.2020].
- <https://arxiv.org/abs/1903.10676> [20.1.2020].
- <https://github.com/kermitt2/delft> [20.1.2020].
- <https://www.tensorflow.org/>, <https://pytorch.org/> [20.1.2020].

Stefan Geißler

Kairntech Deutschland UG
Schützenhausstr 28
69151 Neckargemünd
stefan.geissler@kairntech.com

Stefan Geißler ist Mitgründer von Kairntech mit Sitz in Grenoble, Frankreich. Nach dem Studium der Computerlinguistik war er für einige Jahre am Wissenschaftlichen Zentrum der IBM in Heidelberg beschäftigt. 2000 gründete er mit Kollegen die Softwarefirma TEMIS, als deren deutscher Geschäftsführer er über 18 Jahre hinweg den industriellen Einsatz von NLP vorantrieb.

¹⁵ Vgl. Marcus, Gary und Ernest Davis: „Rebooting AI“, Pantheon, 2019, ISBN 9781524748258.