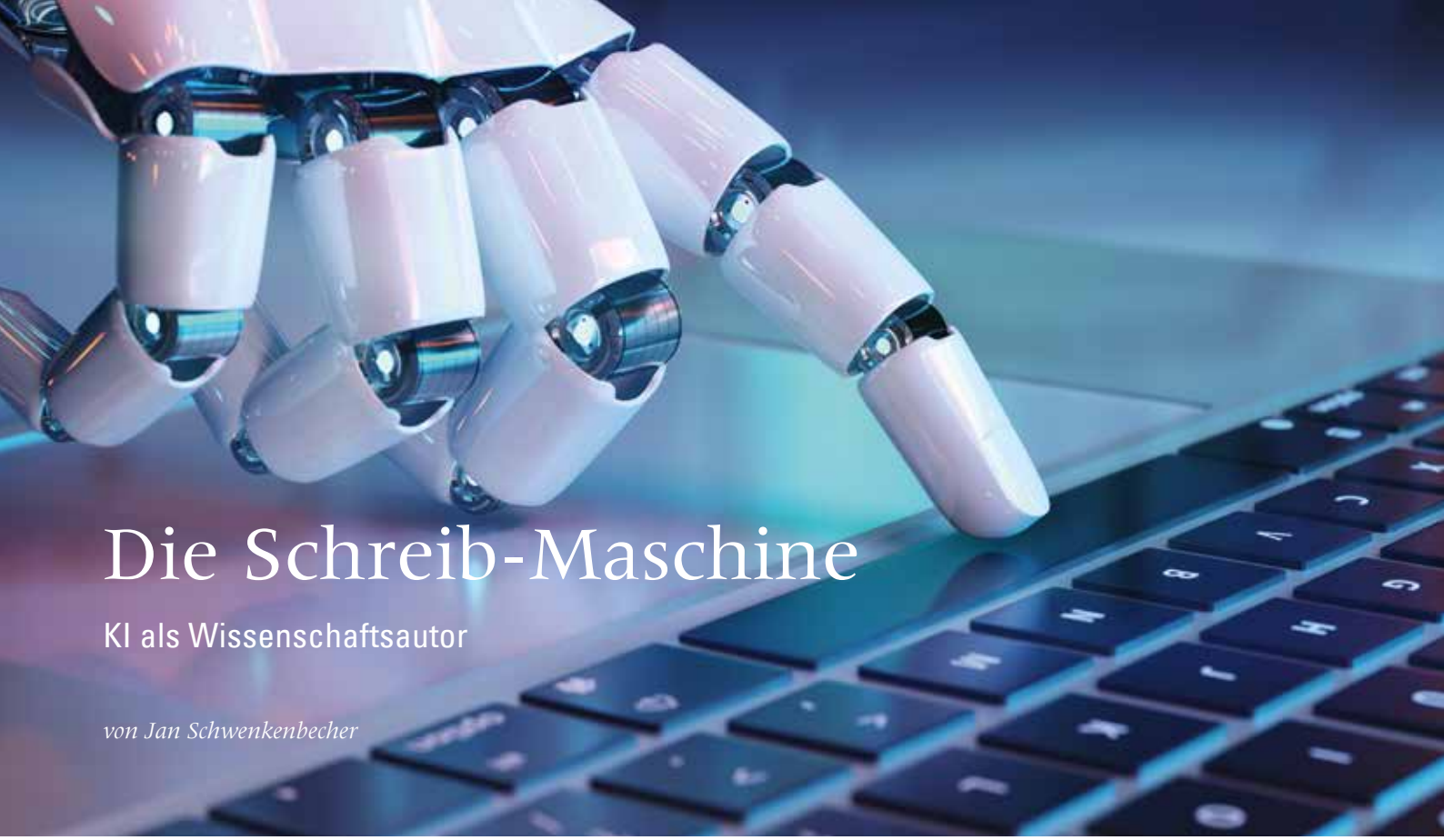


Die Schreib-Maschine

KI als Wissenschaftsautor

von Jan Schwenkenbecher

Der Frankfurter Computerlinguist Christian Chiarcos hat mit seinem Team eine Software entwickelt. Diese Software hat das erste maschinengenerierte Wissenschaftsbuch verfasst.

Auf den ersten Blick erscheint das Buch »Lithium-Ion Batteries« ein ganz gewöhnliches Buch zu sein: Es stammt aus der Feder von »Writer, B.« und ist im Frühjahr 2019 erschienen. Veröffentlicht hat es der Wissenschaftsverlag Springer Nature. Es hat einen Titel, einen Untertitel, es hat vier Kapitel und viele Unterkapitel. Es gibt ein Vorwort, ein Literaturverzeichnis und auch, dass die Überschriften eher hölzern klingen, verwirrt den Leser nicht unbedingt, schließlich handelt es sich um ein Fachbuch. Trotzdem war das Erscheinen des Buchs Springer Nature eine eigene Pressemitteilung wert, und das liegt weniger am Inhalt als vielmehr an »Writer, B.«, dem Autor.

Das »B« steht für Beta und Beta Writer ist nicht der Name eines Batterie-Forschers mit bei der kindlichen Namensgebung etwas zur Exzentrik neigenden Eltern. Beta Writer ist kein Chemiker, auch kein Forscher, Beta Writer ist nicht mal ein Mensch. Die Aufklärung steht auf Seite vier des Buchs: »This book was machine-generated.« Das Buch »Lithium-Ion Batteries« ist das erste von einer künstlichen Intelligenz verfasste Buch, das bei Springer Nature erschienen ist. Beta Writer ist die Software, die es geschrieben hat. Und so gesehen hat Beta Writer dann doch Eltern: ein Forschungsteam, angeführt von Christian Chiarcos, Professor für Angewandte Computerlinguistik und Leiter der Arbeitsgruppe Angewandte Computerlinguistik (ACoLi) der Goethe-Universität Frankfurt, und von Niko Schenk, einem Postdoktoranden seiner Gruppe. Chiarcos wird später über den Beta Writer sagen: »Der Beta Writer ist die Bezeichnung eines Algorithmus, den wir hier auf Basis bestehender eigener Arbeiten und anderer Arbeiten aus der Community erstellt haben, um Bücher zu generieren. Der Plan war, mit ihm das erste maschinengenerierte Wissenschaftsbuch zu erzeugen. Und das haben wir geschafft.«

Auswahl der Quellen

Chiarcos ist auch derjenige, den man fragen muss, wie genau der Beta Writer nun eigentlich das Buch geschrieben hat. Man findet ihn im Frankfurter Stadtteil Bockenheim, am Institut für Informatik. Chiarcos – Strickpulli, Jeans, Brille, braune Haare, die in einen braunen Vollbart übergehen – zieht erst noch einen großen Kaffee aus dem Automaten der Büroküche, dann nimmt er in einem von vier tiefen Sesseln Platz. Ein großes Bücherregal füllt eine Seitenwand, zwei mit unzähligen Formeln beschriebene Tafeln eine andere, auf dem Schreibtisch Dokumentenstapel. Chiarcos nimmt einen Stift in die Hand, mit dem er in die Luft malt, wenn er erklärt, welche vier Schritte er und seine Kollegen auf dem Weg zum KI-Buch gegangen sind.

Maschinen haben das Zeug zum Buchautor und können in Wissenschaftsbüchern einen Literaturüberblick geben.

You can read an English translation of this article online at:
www.aktuelles.uni-frankfurt.de/forschung-frankfurt-englisch

Lithium-Ionen-Batterien sind das Thema von »B. Writers« Buch.

»Der erste Schritt ist das Pre-Processing und man beginnt damit, dass man eine Sammlung von möglichen Quellen aufbaut«, erklärt Christian Chiarcos, anhand welcher Basis das Programm den Inhalt des Buches verfasst hat. Das können PDFs sein oder Word- oder XML-Dokumente. »Diese Quellen haben wir dann nach bestimmten Schlüsselwörtern gefiltert, die wir von Fach-Experten bekommen haben«, sagt Chiarcos, »so haben wir ausgewählt, welche wissenschaftlichen Arbeiten für das Buch relevant waren«. Aus diesen Dokumenten zogen die Forscher den Text heraus, was gar nicht so einfach war, da sich zwischen Wörtern und Satzzeichen jede Menge chemischer Formeln fanden. Aber diese Herausforderung wurde gemeistert und am Ende blieb eine Text-Sammlung aus 1086 Publikationen, alle in englischer Sprache geschrieben und aus der Springer-Nature-Bibliothek.

»B. Writer« schreibt

Im zweiten Schritt setzten die Forscher verschiedene Verfahren ein, um aus dieser Textsammlung eine Struktur für das neue Buch zu gewinnen: die Struktur-Generierung. »Für alle Dokumente haben wir dabei ihre relative Ähnlichkeit zueinander ermittelt«, sagt der Computerlinguist. Wobei sich Ähnlichkeit darauf beziehe, wie ähnlich sich die jeweiligen Texte seien. »Die ähnlichsten werden dann so lange miteinander gruppiert, dass man eine Baumstruktur erhält.« Was wenig Ähnlichkeit aufweise, falle heraus, der Nutzer könne angeben, wie viele Kapitel, Abschnitte und Unterabschnitte er letztlich haben mag und auch, wie viel Text dem Beta Writer in jedem Unterabschnitt zur Verfügung steht, um damit die jeweilige Publikation zusammenzufassen.

»Die eigentliche Text-Generierung, der dritte Schritt, besteht dann daraus, dass man innerhalb eines Textes identifiziert, was die wichtigsten Äußerungen sind«, erklärt Chiarcos. Dazu probierte er mit seinen Kollegen verschiedene Verfahren aus. Eine klassische graphenbasierte Technik, ein moderneres neuronales Modell, am Ende setzten sie die Verfahren parallel ein. In verschiedenen Durchgängen testeten sie verschiedene Gewichtungen und schauten, welches Ergebnis den Experten vom Fach am besten gefiel. Die Experten vom Fach, das waren Chemie- und Batterie-Experten von Springer Nature. Mehrfach legten Chiarcos und seine Kollegen den Experten verschiedene Varianten von Zwischenergebnissen dessen vor, was der Beta Writer bis dahin so zusammengestellt hatte. Die Fachleute bewerteten Inhalt und Stil – wobei sie, wie man beim Lesen des Buches unschwer erkennen kann, fachliche Genauigkeit stärker gewichteten als eine schöne Sprache.

Sätze werden umformuliert

Anhand dieses Feedbacks gewichteten Chiarcos und sein Team denn auch die Verfahren je nach Einsatzort im Buch durchaus unterschiedlich: Die Einleitungstexte jedes Kapitels, die der Beta Writer aus allen darin enthaltenen Publikationen zusammengeschrieben hat, haben eine bestimmte Gewichtung. Die Unterabschnitte, in denen jeweils nur eine einzige Publikation zusammengefasst ist, haben eine andere. Auch für die Kapitelabschnitte »Zusammenfassung« und »verwandte Forschung« gewichteten die Forscher ihre Verfahren wieder neu.

Zur Person

Christian Chiarcos, Jahrgang 1977, Diplom-Informatiker und Sprachwissenschaftler, promovierte 2010 an der Universität Potsdam zum Thema Computerlinguistik. 2012 organisierte er den ersten Workshop zu »Linked Data in Linguistics« in Frankfurt, danach forschte er als Gastwissenschaftler am Information Sciences Institute der University of Southern California in den USA. 2013 erhielt er den Ruf auf die Juniorprofessur für Angewandte Computerlinguistik im Fachbereich Informatik und Mathematik der Goethe-Universität. Begleitend zur Juniorprofessur leitet er seit 2015 die Nachwuchsgruppe »Linked Open Dictionaries«, die vom Bundesforschungsministerium gefördert wird.

chiarcos@informatik.uni-frankfurt.de

Der Text werde dabei wie folgt erstellt: »Wir nehmen einen kompletten Satz«, sagt Christian Chiarcos, »wir eliminieren eventuell Teile davon, wir ersetzen andere Teile davon, wir stellen ihn auf Basis der syntaktischen Analyse um«. Sei der Satz, der dabei herauskommt, dann ausreichend verschieden vom ursprünglichen Satz, werde er nicht als Zitat gekennzeichnet. Die Autoren der Ausgangssätze brauchen sich trotzdem nicht zu sorgen, dass sie da plagiiert werden. Selbst wenn der neue Satz nicht als wörtliches Zitat ins Buch geschrieben werde, stehe immer die entsprechende Fußnote mit der Quellenangabe dahinter.

Auch kritische Stimmen

Schließlich stand im letzten Schritt für Chiarcos und sein Team noch das Post-Processing an. Sie trugen alle Referenzen im Literaturverzeichnis zusammen, fügten die zuvor für die Verarbeitung durch Platzhalter ersetzten chemischen Formeln wieder ein, brachten das Dokument in ein für Springer Nature lesbares Format und übergaben es an den Wissenschaftsverlag.

Und dort scheinen die Verantwortlichen dem Debütroman des Beta Writers die besten Kritiken auszustellen. Henning Schoenenberger, Director Product Data & Metadata Management bei Springer Nature, hat neben Christian Chiarcos und dessen Arbeitskollegen Niko Schenk einen Teil der Einleitung des Buches verfasst und dabei an Lob und Pathos nicht gespart: »Dieses Buch über Lithium-Ionen-Batterien hat das Potenzial, eine neue Ära des wissenschaftlichen Publizierens zu einzuleiten«, so Schoenenberger. Ob das so kommt, das wird erst die Zukunft zeigen. Nach nun etwa einem Jahr wurde das Buch 14-mal zitiert und 357 000-mal heruntergeladen. Es ist allerdings auch kostenlos verfügbar.

Die Downloadzahlen sollten allerdings nicht darüber hinwegtäuschen, dass es am Projekt auch Kritik gab. »Das Feedback, das uns erreicht hat, war zwar weitgehend positiv«, sagt Chiarcos. Es habe aber auch ein paar sehr kritische Stimmen zur Frage nach der sozialen, politischen Verantwortung gegeben. »Leute haben die wissenschaftliche Verantwortung betont und nachgefragt, ob das System nicht ein verzerrtes Bild eines Fachgebiets erzeuge, einen Bias.«

Feinere Sprache, schönere Überschriften

Tatsächlich werden die aufgenommenen Publikationen ja anhand ihrer Ähnlichkeit untereinander ausgewählt. Wenn nun schon diese Ursprungsdaten die Wirklichkeit verzerrt abbilden – etwa, weil jemand eine bestimmte Forschungsrichtung oder eine bestimmte Forschungsgruppe umfangreich finanziert und auf diesem Teilgebiet nun besonders viele Publikationen vorliegen –, dann übernimmt das System diese Verzerrung und verstärkt den Bias. »Unser System erzeugt eine solche Verzerrung zwar nicht«, so Chiarcos, »aber es gibt keinen Weg, das automatisch zu kompensieren. Das geht nur, wenn sich ein Experte vom Fach dransetzt und die Literatur manuell sichtet«.

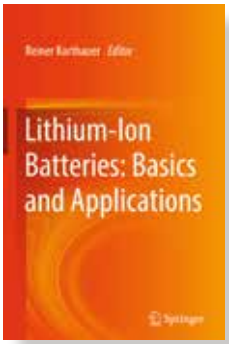
Daneben existieren noch zahlreiche weitere Punkte, die Christian Chiarcos mit seinen Kollegen gerne optimieren würde. Eine feinere Sprache. Schönere Überschriften. Stärkere Kohärenz. Darüber hinaus gibt es neben Lithium-Ionen-Batterien ja auch noch andere Forschungsfelder in der Wissenschaft, in denen der Beta Writer ebenfalls den ein oder anderen Sammelband zusammenstellen könnte.

Das nun erschienene Buch hat die Frage beantwortet, ob künstliche Intelligenz wissenschaftliche Bücher verfassen kann. Sie kann. Die sich nun stellende Folgefrage lautet: in welcher Rolle der Beta Writer – oder ihm ähnliche Algorithmen – in die Bibliotheken der Wissenschaftsverlage einziehen. Wird hier und da mal ein Review entstehen? Oder sind bald alle Sachbuchautoren des Landes ihre Jobs los?

Für den persönlichen Literaturüberblick

Die eigentliche Stärke des Beta Writers liegt gar nicht darin, dass er ein wissenschaftliches Buch geschrieben hat. Sie liegt darin, dass er ein wissenschaftliches Buch über ein x-beliebiges Forschungsthema geschrieben hat und die Anwender – in dem Fall Christian Chiarcos und Kollegen – dem Programm noch vorgeben konnten, wie viele Kapitel sie gerne hätten und wie lang diese sein sollen. Vielleicht könnte die Hauptarbeit des Beta Writers damit eine ganz andere werden, als Bücher zu schreiben. Schließlich ist es eine Software, die automatisch einen ganz individuellen Literaturüberblick schaffen kann. Den brauchen etwa Forscher, wenn sie sich einem neuen Thema widmen, den brauchen aber auch Doktoranden beim Verfassen der Abschlussarbeit.

»Tatsächlich ist das, was ich perspektivisch für die wahrscheinlichste Anwendung dieser Technologie halte«, sagt auch Christian Chiarcos. Er glaube, dass man die Software gar nicht so sehr als ein Generierungs-Tool verwenden werde. »Sondern eher als ein Werkzeug, das einem Menschen dabei hilft, effektiver Bücher zu schreiben.« ●



»B. Writer«
Reiner Korthauer, Hg.
Lithium-Ion Batteries: Basics and Applications

Das erste Buch der KI »B. Writer«. Kostenloser Download unter https://tinyurl.com/BWriterBattery

Aus aktuellem Anlass ein weiteres Buch von »B. Writer«, ein Literaturüberblick zu SARS-CoV-2 und COVID-19: https://tinyurl.com/BWriterCovid

Der Autor

Jan Schwenkenbecher ist freier Wissenschaftsjournalist und lebt im Rhein-Main-Gebiet. Er hat in Gießen und Mainz Psychologie studiert und danach im Volontariat bei der Süddeutschen Zeitung das journalistische Handwerk gelernt.

jan.schwenkenbecher@posteo.de