



Zusammenfassung

Dieses Kapitel adressiert KI-Systeme, die sich mit der Verarbeitung, dem Verständnis und der Veränderung von Text in seinen verschiedenen Formen beschäftigen. Ziel ist es, ein besseres Verständnis für diese Art Systeme zu ermöglichen, die Grundlage der eingesetzten Methoden darzulegen und so die Beurteilung der Leistungsfähigkeit textverarbeitender KI zu unterstützen. Dazu werden grundlegende Technologien und Konzepte vorgestellt, sowie ein Einblick in konkrete Anwendungsfälle wie bspw. Einfache Sprache gegeben.

10.1 Einleitung

Vor mehr als viertausend Jahren entstand das erste richtige Alphabet – in Zenträgypten, wo verschiedene Hieroglyphen die Laute der Sprache repräsentierten. Knapp dreitausendfünfhundert Jahre später, 1440, wurde der Buchdruck mit beweglichen Lettern von Gutenberg in Mainz erfunden. Die Geschichte der Menschheit ist von geschriebener Kommunikation geprägt, Kultur wäre ohne sie undenkbar. Daher scheint es naheliegend, dass Texte zu verstehen und zu produzieren auch eine zentrale Aufgabe von KI-Systemen geworden ist.

Die automatisierte Verarbeitung von Textdokumenten stellt gerade für die Verwaltung eine der wichtigsten Neuerung dar, die durch KI erreicht werden können. Aber es ist keine leichte Aufgabe – Ein „nicht“ in einem Satz dreht auf einmal den Sinn des Satzes um. Ein „möglich“ verändert, ob eine Information als gegeben gesichert werden darf oder nicht. Rechtschreibfehler können in jedem Text auftreten – trotzdem sollte und kann man als Mensch verstehen, was in dem Satz gemeint ist. Wir wollen auch effizient sprechen, benutzen Pronomen anstatt

immer die Nomen. Auf welches Wort bezieht sich aber ein Demonstrativpronomen wie „diese“ – das ist schon für uns Menschen beim Lesen manchmal schwer. Wie kann eine KI das erreichen?

Dieses Kapitel beschäftigt sich mit der Frage, was KI-Textverarbeitung ist, wie sie funktioniert und wo sie eingesetzt werden kann. Es werden grundlegende Technologien für die Textverarbeitung vorgestellt. Es wird ein spezifisches Beispielprogramm gezeigt, das versucht, die Semantik in rechtlichen Texten zu erfassen und mit anderen zu vergleichen.

Relevant sind die Rahmenbedingungen, denn die Methoden zur digitalen Textverarbeitung sind inzwischen sehr leistungsfähig. Anfragen, zum Beispiel in Suchmaschinen, werden häufig hervorragend verarbeitet. Gerade bei Texten ist die Frage nach dem Digitalisierungsgrad sehr wichtig. Welche der Dokumente, die einbezogen werden sollen, liegen digital vor? Kommen Dokumente nur per Post? Stehen Texte, mit denen ein System trainiert werden kann, so wie in der Kap. 3 dargestellt, zur Verfügung? KI-Textverarbeitung kann nur gelingen, wenn die Voraussetzungen erfüllt sind.

10.2 Grundlagen der Textverarbeitung

Nachdem nun Fallbeispiele etabliert worden sind, geht es in einem ersten Schritt darum, ein grundlegendes Verständnis dafür zu entwickeln, was „Textverarbeitung“ im Kontext von Künstlicher Intelligenz bzw. Machine Learning überhaupt bedeutet. Vereinfacht gesprochen umfasst Textverarbeitung alles, was mit der Kategorisierung von Text zu organisierten Gruppen zu tun hat. Technologien wie Natural Language Processing werden dabei genutzt, um Texte nach bestimmten Bedingungen zu klassifizieren, zu unterteilen oder auszuwerten, sodass neue Erkenntnisse oder Ordnungsstrukturen daraus generiert werden können (vgl. Russel & Norvig, 2012, S. 1001).

Typische Anwendungsfälle, bei denen Textverarbeitung zum Einsatz kommt, sind bspw. Online-Publikationen von Blog- oder News-Artikeln, die automatisiert mit einem thematischen Tag versehen werden. Tonalitätsanalysetools, die den „Ton“ eines Zeitungsartikels oder von Konversationen in sozialen Netzwerken bestimmen sollen. Aber auch die Einschätzung von Krisenlagen durch Konversationen über soziale Netzwerke, etwa im Falle einer Naturkatastrophe oder eines Terroranschlags.

Bevor wir nun auf konkrete Techniken eingehen, die bei der Textverarbeitung genutzt werden, muss zunächst dargestellt werden, was bei Prozessen zur Textverarbeitung überhaupt mit dem Text gemacht wird:

- Eine der simpelsten Möglichkeiten ist die Anwendung einfacher statistischer Methoden. So kann die Häufigkeit von bestimmten Worten gezählt und statistisch erfasst werden, um den Text zu kategorisieren. Die Häufigkeit, mit der z. B. ein bestimmter Paragraph oder ein Projektname in einer E-Mail auftaucht, kann so Auskunft darüber geben, wie der Ursprungstext thematisch einzuordnen ist.
- Des Weiteren kann auch das bereits aus Kap. 4 bekannte Pattern Matching genutzt werden, um Texte nach ihren inhärenten Mustern und Strukturen zu klassifizieren. Die Klassifizierung kann beispielsweise darin bestehen, bestimmte Formulierungsfloskeln oder Zitate aus Gesetzestexten zu erkennen. Gleichzeitig können aber auch komplexe Strukturen aufgedeckt werden, die für die jeweiligen Rezipienten ohne die Hilfe des Algorithmus nicht oder nur mit Schwierigkeiten erkennbar gewesen wären. Ein mögliches Beispiel sind Texte, die ein ähnliches Level an komplexer Sprache verwenden o. ä.
- Eine weitere, exemplarische Möglichkeit für die Untersuchung von Texten im Rahmen der Textverarbeitung, die hier vorgestellt werden soll, ist die Möglichkeit, den Text auf Grundlage von verschiedenen, vorher festgelegten Regeln zu untersuchen. Bspw. sollen alle Texte, die Paragraphen enthalten, auf eine bestimmte Art klassifiziert werden oder es werden alle Texte, die eine bestimmte Kombination von Worthäufigkeiten und Merkmalen aufweisen, in eine definierte Gruppe eingeordnet.

Da nun die grundlegenden Prozesse bekannt sind, mit denen Textverarbeitung durchgeführt wird, stellt sich als Nächstes die Frage, auf welche Art und Weise sich unterschiedliche Anwendungen zur Textverarbeitung bei der Umsetzung dieser Methoden unterscheiden. Eine der einfachsten, aber nicht minder relevanten Möglichkeiten, die Fähigkeiten einer Anwendung für die Textverarbeitung einzuschätzen, ist, die Größe der untersuchten Textsegmente zu betrachten, also vereinfacht gesagt, wie viel Text kann vom Algorithmus parallel überprüft und verarbeitet werden. Weitere Unterscheidungsmöglichkeiten ergeben sich aus den Fähigkeiten des Algorithmus semantische Unterschiede (Unterschiede auf der Bedeutungsebene des Textes) zu erkennen und zu verarbeiten. Praktische Beispiele für diese semantische Differenzierung sind z. B. die Fähigkeit, Verneinungen als solche zu erkennen und zu bewerten oder aber auch wie gut

der Algorithmus darin ist, die Relevanz verschiedener Textstellen füreinander zu klassifizieren.

Ausgehend von den vorgestellten Methoden und Differenzierungsmöglichkeiten sollen abschließend noch drei Verfahren vorgestellt werden, auf deren Grundlage Algorithmen zur Textverarbeitung operieren können (vgl. Chilakapati, 2019):

Bag of Words

Eine der grundlegendsten Verfahren ist dabei der sogenannte Bag of Words (BoW). Hier wird die Häufigkeit, mit der Worte in einer zu untersuchenden Textmenge auftreten, genutzt, um Classifier zu trainieren, die schlussendlich Aussagen über andere Texte ermöglichen. Stellen wir uns exemplarisch vor, dass wir einen Algorithmus nutzen wollen, um Begrüßungen in E-Mails danach zu klassifizieren, ob sie eher formell oder eher informell von ihrer Ansprache her sind. Der erste Schritt ist das Bereitstellen von Trainingsdaten, mit denen der Algorithmus trainiert werden kann. Sobald wir eine Grundmenge an Beispiel-E-Mails gesammelt haben, müssen diese von uns mit der Dimension, die wir untersuchen wollen, klassifiziert werden. Hier, ob sie formell = 1 oder informell = 0 sind. Diese Trainingsdaten werden dem Algorithmus als Grundlage zur Verfügung gestellt, der, wie bereits beschrieben, schlicht die Häufigkeit der Worte erfasst, die in den Beispieldaten vorkommen. Für unser Beispiel der Sprachformalisierung würde das bedeuten, dass z. B. in 100 als formell beschriebenen E-Mails 95 Mal das Wort „geehrte“ vorkommt, während bei den informellen „Hallo“ die höchste Häufigkeit hat. Auf Grundlage dieser Häufigkeiten und in Abhängigkeit der jeweiligen Zuordnung (formell vs. informell) kann der Algorithmus nun Vorhersagen darüber generieren, ob neue Texte, die ihm zur Verfügung gestellt werden, eher formell oder informell formuliert worden sind.

Word2Vec

Das zweite Verfahren, das betrachtet werden soll, ist die sogenannte Word2Vec (W2V) Methode. Bei W2V geht es darum, aus Worten Vektoren zu errechnen, deren mathematische Nähe zueinander die semantische Nähe der Worte abbildet. Mit einem ausreichend großen Datensatz ermöglichen es die unter dem W2V-Begriff gesammelten Verfahren, in Abhängigkeit des Auftretens der Worte in den Trainingsdaten, recht präzise Aussagen über die Bedeutung eines Wortes zu formulieren. Betrachten wir die Worte Hund, Welpen, Katze und Kätzchen unter der vorgestellten Prämisse, dann würden wir eine größere Nähe zwischen Hund und Welpen bzw. Katze und Kätzchen feststellen, während sich alle in gleicher Distanz zu Mensch befinden könnten. Das Ergebnis von Word2Vec sind nicht nur einzelne Vektoren, sondern ein komplexes Netz, das die Beziehungen der Worte zueinander in Abhängigkeit ihrer

Kosinusähnlichkeit abbildet. Dies ermöglicht es uns, die Worte in Abhängigkeit ihrer Ähnlichkeit zueinander von 0 = keine Ähnlichkeit (90°) bis 1 = volle Ähnlichkeit (0°) darzustellen. Diese Vektorenetze können dann genutzt werden, um weitere Algorithmen mit Informationen zu versorgen. Im Alltag finden wir sie vor allem im Hintergrund von Suchmaschinen oder den diversen E-Commerce Seiten, die wir immer wieder nutzen.

Bidirectional Encoder Representation From Transformers (BERT)

Das letzte Verfahren, das im Rahmen dieses Segments beleuchtet werden soll, ist die Bidirectional Encoder Representation From Transformers oder kurz BERT. BERT ist das modernste der vorgestellten Verfahren und adressiert einige der Schwachstellen zuvor präsentierter Modelle. Die Wichtigsten sind dabei:

1. Die Erkenntnis, dass Bag of Words semantikagnostisch agieren d. h., dass sie nur über die Häufigkeit der zur Verfügung gestellten Worte klassifizieren. Weder bilden sie die Bedeutung des Textes adäquat ab, noch ist gewährleistet, dass „richtige“ Worte genutzt wurden, um den Algorithmus anzulernen.
2. Die Tatsache, dass es lediglich einen Vektor pro Wort gibt, was die sinnvolle Einbindung von Dimensionen wie „Kontext, in dem das Wort verwendet worden ist“ oder „Position, an dem es im Satz steht“ schwierig abbildbar macht. Dies verursacht vor allem dann Probleme, wenn Synonyme/Antonyme die Positionierung des Wortes beeinflussen würden, einem Wort mehrere Bedeutungen zukommen könnte (bspw. um-fahren = ausweichen vs. umfahren = etwas mit dem Vehikel umstoßen/überrollen) oder es sowohl als Verb, aber auch nominalisiert auftreten kann (laufen vs. das Laufen).

BERT versucht dieses Problem aufzulösen oder zumindest zu reduzieren, indem es Wort-Vektoren generiert, die sich abhängig von ihrem aktuellen Ort (in der Word Map) und den anderen Wort-Vektoren in ihrer Umgebung anpassen. Es wird also versucht, sowohl Kontext als auch Semantik mit in die Urteilsfindung des Algorithmus einzubeziehen. Dies geschieht mithilfe eines Aufmerksamkeitsmechanismus: Während zuvor nur einzelne Worte einen Vektor gebildet haben, werden nun die Beziehungen jedes einzelnen Wortes in einem Satz zu allen anderen bekannten Worten als ein sogenannter Kontext-Vektor dargestellt. Hier kommt die „Aufmerksamkeit“ ins Spiel. Mit Self-attention, also den Fokus auf ein Wort selbst oder Inter-Attention, den Fokus auf die Beziehung zwischen Worten, wird es möglich, eine Beschreibungskategorie zu entwickeln, mit deren Hilfe Bedeutung verlässlicher abgebildet werden kann. Während wir meistens davon ausgehen, dass ein Satz

nur eine mögliche Auslegung hat und nicht weiter über die einzelnen Worte nachdenken, wird schnell klar, dass einzelne Worte eine engere semantische Beziehung zueinander und somit auch zum gesamten „Wortraum“ haben als andere im gleichen Satz.

Im Vergleich mit den anderen vorgestellten Verfahren kann dies als Novum bezeichnet werden, da es nicht nur akkuratere Ergebnisse ermöglicht, sondern auch zum ersten Mal eine Form von Aufmerksamkeit nutzt, die es dem Algorithmus erlaubt, Sätze als multidimensionale Objekte zu verstehen, deren Bedeutung zwar abbildbar ist, sich aber in Abhängigkeit verschiedener Kontexte ständig verändern kann.

Übung

1. Bei welcher Methodik zur Textverarbeitung werden im Wesentlichen die Distanzen einzelner Begriffe genutzt, um ein semantisches Netz zu konstruieren?
 - a) Word2Vec
 - b) Bag of Words
 - c) BERT
 - 5) Classifier-VII
2. Was bedeutet „semantikagnostisch“ im Bezug auf die Textanalyse?
 - a) Der Algorithmus arbeitet nur über die Häufigkeit von Begriffen, nicht deren Inhalt
 - b) Der Algorithmus kann mit verschiedenen Inhalten arbeiten, nicht nur mit bestimmten
 - c) Der Algorithmus lernt die Semantik jedes Textes einzeln und ist nicht vorher schon trainiert
 - d) Der Algorithmus beachtet ausschließlich Semantik von Sätzen, nicht aber deren Syntax
3. Welche Stärke hat BERT gegenüber anderen, vektor-basierten Verfahren ?
 - a) BERT wurde mit einem deutlich kinderfreundlicheren Datensatz trainiert
 - b) BERT wird durch eine unabhängige Expertenkommission kontrolliert
 - c) BERT kann mit unterschiedlichen Verwendungen für ein- und dasselbe Wort besser umgehen
 - d) BERT erlaubt es die semantische Gesamtbedeutung eines Textes inhärent mit anderen Texten zu vergleichen ohne jedes Wort einzulesen

10.3 Natural-Language-Processing-Bestandteile: Intent, Entity, Kontext und Dialogue Management

Das Verstehen von Sprache ist eine enorm herausfordernde Aufgabe, insbesondere dann, wenn ein System Aussagen oder Aufforderungen im Rahmen eines Dialogs erkennen und korrekt bearbeiten soll. In diesem Kapitel soll es daher etwas detaillierter um die verschiedenen Komponenten einer Nachricht gehen, die z. B. von einem Chat-Bot identifiziert und berücksichtigt werden müssen (vgl. dazu Jain et al., 2018).

Schauen wir uns dafür zunächst verschiedene Beispiele in den beiden Programmen SchreibFix und Memoriali an:

1. Eine Nutzerin fordert SchreibFix auf, ihre Mail am kommenden Tag, um 15.00 Uhr, zu versenden.
2. Ein Sachbearbeiter fragt Memoriali: „Welcher Gutachter wohnt in der Nähe?“.
3. Ein Bürger fragt das Memoriali-System, wie die Vermessung von Fenstern funktioniert.
4. Eine Kollegin ist verärgert über SchreibFix und fragt es, warum es so unfähig sei.

Alle diese Ausdrücke werden also an ein System weitergeleitet. Die Bearbeitung von Anfragen, Aussagen und Ausdrücken übernimmt dabei das sogenannte „Dialog Management“. Dieses System hat die Aufgabe, die sogenannte „Dialog Handlung“ zu identifizieren. Damit ist gemeint, welche Funktion eine bestimmte Aussage im Rahmen des Dialogs erfüllt und welche Antwort diese korrekt bedienen könnte. So stellt „15 km“ z. B. keine Antwort für das zweite gegebene Beispiel dar.

Das Dialog Management hat daher die Aufgabe, verschiedene Aspekte des Dialogs und der spezifischen Dialog Handlung zu identifizieren, um eine passende Funktion zuzuschreiben und zu antworten. Dazu gehören im Wesentlichen folgende Aspekte: die Entities einer Aussage bzw. des Dialogs, der Kontext des Dialogs sowie der durch einen Ausdruck dargestellte Intent.

Entities sind dabei die relevanten Elemente bzw. Objekte in der Aussage eines Nutzers. In Beispiel 3 stellen zum Beispiel die Fenster eine Entity dar, die für die Aussage des Nutzers relevant ist. Die Vermessung stellt ebenso eine Entity dar, die notwendig ist, um den Ausdruck des Satzes zu verstehen. Das System kann diese beiden Entities erkennen und sie in der Bearbeitung der Anfrage einsetzen. Bei der Entwicklung eigener Systeme ist es wichtig, dass klar ist, welche Entities es gibt bzw. welche das System beherrschen soll. Würde in Beispiel 3

nun nicht nach Fenstern, sondern z. B. Türen, gefragt werden, müsste die Entity Tür genauso bekannt sein (würde aber vermutlich zu einem anderen Ergebnis führen). Würde nach z. B. Teppichen gefragt werden, würde Memoriali diese Entity vermutlich nicht bearbeiten können. Es kann daher auch hilfreich sein, den Nutzenden Beispiele für Entities zu zeigen.

Aus der Entity-Übersicht alleine wird jedoch noch keine Dialog Handlung, ansonsten würde „Vermessung Fenster“ ja bereits reichen. Es bleibt aber offen, welche Intention (also welchen Intent) die Person mit diesen Entities verfolgt. Soll eine Vermessung eingereicht werden? Soll eine Vermessung angefordert oder beauftragt werden? In diesem Fall ist der Intent eine Anleitung oder Erklärung. Der Intent wäre also zum Beispiel „Gib mir eine Erklärung“ – die betroffenen Entities wären „Fenster“ und „Vermessung“. Diese Informationen reichen dem Dialog Management, um die Funktion der Aussage zu erkennen und – sofern möglich – eine passende Antwort abzugeben.

Hier zeigt sich auch, wieso eine gebrauchstaugliche Entwicklung solcher Dialog-Systeme nicht immer einfach ist: die Entity „Vermessung“ könnte auch durch „Berechnung“, „Ausmaße“, „Größenangabe“, „Flächenmessung“ oder andere Begriffe adressiert werden. Sind diese im System jedoch nicht hinterlegt, sondern nur der Begriff „Vermessung“ als Entity bekannt, kommt es vermutlich zu einem Fehler. Moderne Sprachassistenten stehen häufig vor diesem Problem und greifen dafür z. B. auf große, semantische Datenbanken zurück.

Jedoch lassen sich auch mit Entity und Intent nicht alle Anfragen zufriedenstellend beantworten. Ein Beispiel dafür stellt die dritte Aussage dar. Um nach der erfolgreich erkannten Funktion der Aussage eine passende Antwort zu liefern, benötigt das Dialog Management nämlich die Information, wo die Frage gestellt worden ist. Dies wird mit dem Kontext des Dialogs ausgedrückt. Alle Meta-Informationen, die für das Dialog Management relevant sind, finden sich hier wieder. Dies kann auch die Identität des Fragestellenden, die Uhrzeit oder die Länge der Unterhaltung beinhalten.

Liegen Kontext, Entities und ein erfolgreich erkannter Intent vor, sind die wichtigsten Bestandteile zur Erkennung einer Anfrage vorhanden.

Übung

1. Eine Nutzerin fragt: „Welche Kosten verursacht eine Neuausstellung des Personalausweises?“ – welche der folgenden Aussagen dazu ist korrekt?
 - a) Der Kontext könnte sein: „Neuausstellung Personalausweis,“
 - b) Bei „Kosten“ handelt es sich nicht um eine Entity
 - c) Die Funktion der Frage ist „Kostenauskunft,“

- d) Diese Aussage beinhaltet keinen Intent
2. Welche der folgenden Antworten ist kein geeignetes Beispiel für den „Kontext“ einer Anfrage?
- a) Person befindet sich an der Infotheke des Rathauses
 - b) Person hat bereits vier fehlgeschlagene Anfragen ausgeführt
 - c) Person fragt nach einer Wegbeschreibung
 - d) Alle oben genannten sind Kontextbeschreibungen
3. Was ist die Aufgabe des Dialog Managements?
- a) Relevante Elemente der Unterhaltung identifizieren und Antwort auswählen
 - b) Überwachung von Antwortlatenz und –tendenz nach Nutzerpräferenz
 - c) Überprüfung des Wahrheitsgehaltes von Aussagen vor der Antwort
 - d) Aufrechterhaltung personalisierter Informationen zur Präzisierung der Antwort im Text

10.4 Ziele von Textverarbeitung

Da nun die Grundlagen zum Thema der Textverarbeitung etabliert worden sind, kann sich nun den Zielen von Textverarbeitung zugewandt werden. Dazu gilt es zunächst die Anwendungsbereiche klarer zu umreißen.

10.4.1 Anwendungsbereiche von Textverarbeitung

Dabei können wir grundlegende Anwendungsfälle definieren, auf die im weiteren Verlauf genauer eingegangen werden soll:

- **Analyse von Texten:** Hier geht es darum, dass Algorithmen genutzt werden, um Texte zu interpretieren bzw. Informationen aus ihnen zu generieren. Dies kann bspw. in Form einer semantischen Analyse der Fall sein, bei der nach bestimmten Bedeutungen und Inhalten gesucht wird z. B. dem Ton eines Anschreibens oder ob eine bestimmte Klausel im Text enthalten ist. Eine weitere Art der Analyse wäre die Auswertung sonstiger Informationen wie der Häufigkeit bestimmter Worte oder Zeichen.
- **Bearbeitung von Texten:** Bei der Bearbeitung geht es um die Manipulation oder Veränderung von bereits existierenden Textdateien. Anwendungsfälle umfassen die Korrektur von Rechtschreibung und Orthographie in Texten, die Umwandlung in andere Sprachformen (einfache Sprache, Übersetzungen,

etc.) oder aber auch das Schwärzen von Unterlagen in Abhängigkeit ihrer Geheimhaltung.

- **Erstellen von Texten:** Algorithmen können ebenfalls dazu genutzt werden, ganze Texte oder auch einzelne Textsegmente auf Grundlage zuvor definierter Inputs zu generieren. Exemplarisches Beispiel wäre die Generierung des Anschreibens eines Kundenbriefes auf Grundlage von Faktoren wie Ton, Grund des Schreibens (Beispiel SchreibFix).
- **Synthetisierung von Ergebnissen:** Bei der Synthetisierung von Ergebnissen auf Grundlage von Textdateien werden Textverarbeitungsalgorithmen genutzt, um neue Inhalte auf Grundlage des verarbeiteten Textes zu generieren. Dabei ist sie eng mit der Analyse von Texten verbunden. Während „reine“ Analyseinhalte meist Metainformationen über den Text widerspiegeln, dieser also noch für das Verständnis des generierten Outputs gebraucht wird, lassen sich die Ergebnisse von Synthetisierung auch unabhängig ihres Ursprungs verstehen.

10.4.2 Fallbeispiele

Nachdem wir diese Klassifizierung von verschiedenen Anwendungsfällen für Textverarbeitung jetzt verinnerlicht haben, sollen nun ein paar anwendungsnahe Beispiele aus dem Arbeitsalltag vorgestellt werden:

Memoriali

Betrachten wir unser Fallbeispiel Memoriali für die Begutachtung von Altbauten. Hier könnte ein Algorithmus zur Textverarbeitung eingesetzt werden, um aus den Massen an Unterlagen, die für die einzelnen Immobilien zur Verfügung gestellt werden, die wichtigsten Argumente zu extrahieren und für den/die Sachbearbeiter/ in aufzuarbeiten. Die die Immobilie betreuende Person könnte so beispielsweise nur die Stichworte „Energie“ und „Dämmung“ in eine Suchmaske eingeben und der Algorithmus würde nicht nur eine Darstellung aller Dokumente liefern, in denen diese Begriffe vorkommen, sondern auch alle aufzeigen, die damit auf einer Bedeutungsebene assoziiert sind (bspw. in Form von Rechnungen für energetische Maßnahmen).

Einfache Sprache

Ein weiteres Beispiel für die Bearbeitung von Text durch Algorithmen, das in unserem Alltag vielleicht häufiger auftaucht als zunächst offensichtlich, ist die Übersetzung von Texten in einfache Sprache. Viele Einrichtungen im Bereich

der öffentlichen Verwaltung bieten die Inhalte ihrer jeweiligen Websites auch in einfacher Sprache an. Einfache oder Bürger-Sprache (vgl. Bürgernahe Verwaltungssprache, 2002) ist eine vereinfachte Version der Standardsprache, deren Fokus auf der Verständlichkeit des Kommunizierten liegt. Sie zeichnet sich vor allem durch einfachen und kurzen Satzbau sowie die Verwendung von möglichst wenig Fachworten aus. Hier greift die Textverarbeitung ein. Algorithmen können dazu genutzt werden, komplizierte oder fachspezifische Texte zu verarbeiten und eine Version in einfacher Sprache auszugeben. Zusätzlich ist es auch möglich, Texte von Grund auf durch einen Algorithmus in einfacher Sprache schreiben zu lassen wie bspw. dieser Artikel im Guardian, der komplett durch GPT-3 geschrieben wurde: <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>.

SchreibFix

Andere Beispiele für die Erschaffung von ganzen Texten durch einen angelernten Algorithmus ist unser Fallbeispiel SchreibFix. Hier kann das Tool genutzt werden, um auf Grundlage verschiedener Faktoren wie dem gewünschten Ton der E-Mail oder dem bisherigen Verlauf der Konversation, ganze E-Mails zu generieren, die dann durch die Nutzenden sofort verwendet werden können.

Synthetisierte Ergebnisse

Ein weiteres Beispiel ist die Erstellung von synthetisierten Ergebnissen. Diese lassen sich in der Strafverfolgung finden. So können Algorithmen genutzt werden, um die Textdateien, die bei der Beschreibung von gesuchten Personen erstellt werden, zu verarbeiten, sodass aus ihnen heraus neue mediale Formen generiert werden können. Im Falle der Personenbeschreibungen könnte der Algorithmus eingesetzt werden, um aus Informationsteilen wie „Größe“, „Augen- und Haarfarbe“, „Körperbau“ usw. ein Bild zu generieren, das die gesuchte Person zeigt. Ein Prozess, der nicht nur die Erstellung von Fahndungsbildern vereinfacht, sondern es auch ermöglicht, in kurzer Zeit verschiedene Varianten einer gesuchten Person zu generieren.

Übung

1. Was ist kein Ziel von Textverarbeitung?
 - a) Die Manipulation bereits existierender Texte
 - b) Die Überprüfung von Rechtschreibregeln
 - c) Die Analyse des Tonfalls eines Textes
 - d) Die Beschreibung eines Bildes durch Text

2. Welche dieser Aussagen ist falsch
 - a) Durch Textverarbeitung dürfen sich der Inhalt und Sinn eines Textes nicht ändern
 - b) Durch Textverarbeitung können deutlich kürzere oder längere Texte entstehen
 - c) Nicht jede Form der Textverarbeitung ist inhärent erklärbar
 - d) Die Übersetzung von Texten ist eine spezifische Form der Textverarbeitung
3. Welcher Aspekt ist nicht Teil von einfacher Sprache?
 - a) Fokussierung auf simplen Satzbau
 - b) Kurze Sätze ohne viele Einschübe/Nebensätze
 - c) Nur geringe Verwendung von Fachworten
 - d) Große Zeilenabstände für bessere Lesbarkeit

10.5 Fallbeispiel Semantha

In dem vorliegenden Fallbeispiel möchten wir ein existierendes Produkt, welches im Bereich der sprachverarbeitenden KI existiert, genauer unter die Lupe nehmen. Dafür wird zunächst ein Einsatzszenario und die Aufgabe, die das Tool dabei übernehmen kann beschrieben. Im Anschluss wird diskutiert, welche Auswirkungen dies auf Arbeitsprozesse haben kann.

Hinweis: Im vorliegenden Fall geht es um die Nutzung des Tools Semantha von der Firma thingsTHINKING. Die Anwendung der Software wurde für die Diskussion in diesem Buch abgewandelt und entspricht unter Umständen nicht dem (aktuellen) Funktionsumfang des tatsächlichen Produktes.

Cem ist Sachbearbeiter und kümmert sich um alle Anliegen, die mit dem Denkmalschutz in Verbindung stehen. Derzeit ist seine Aufgabe, die Leistungsbeschreibungen eines städtischen Bauherren für Umbaumaßnahmen an einem Gebäude zu prüfen. Er soll feststellen, ob diese mit den Vorschriften übereinstimmen. Dieser Prozess ist mühselig, da das Auftragsbuch enorm umfangreich ist und für verschiedene Etagen, Fenstergrößen, etc. einzelne Beschreibungen des Austauschprozesses angefertigt worden sind. Cem entscheidet, dass eine genaue Prüfung zu viel Zeit in Anspruch nehmen würde. Einerseits hat er selbst genug zu tun, andererseits ist das Projekt schon im Verzug. Er wählt die Leistungsbeschreibung von einem Fenster je Etage und segnet danach die Leistungsübersicht ab.

Mithilfe eines sprachverarbeitenden Tools könnte Cem hier auch anders vorgehen. Hinter Semantha steckt ein Algorithmus, der in der Lage ist, den semantischen Inhalt von Texten zu erfassen, auch wenn diese nicht wortwörtlich

miteinander übereinstimmen. Dies kann hilfreich sein, wenn zum Beispiel unterschiedliche Rechtstexte schnell miteinander verglichen werden müssen. Auch bei der Analyse oder dem Vergleich von Verträgen können Tools zur Verarbeitung semantischer Daten genutzt werden.

Was würde in dem Beispiel genau geschehen? Anstatt den Auftragstext komplett zu lesen, hat Cem sich entschieden, die Korrektheit stichprobenartig zu überprüfen. Er versucht so, seine Arbeitslast zu reduzieren und dennoch ein verlässliches Ergebnis zu erreichen. Die Stichproben wählt er zufällig, hier z. B. in jeder Etage ein Fenster, aus.

Beim Einsatz von Semantha würde sich dieser Arbeitsprozess verändern. Die Sprachverarbeitungssoftware würde jede Leistungsbeschreibung automatisiert mit den entsprechenden Vorschriften für den Denkmalschutz bzw. entsprechende Umbaumaßnahmen abgleichen. Die Stärke eines KI-basierten Ansatzes ist dabei, dass zum Beispiel ein Begriff wie „nicht abschließend“ und „undicht“ im Bezug auf Fenster als gleichwertig betrachtet werden. Dadurch ist es möglich, Texte miteinander zu vergleichen, die zwar unterschiedliche Autoren bzw. Wortschätze haben, aber dasselbe Thema behandeln.

In unserem Beispiel könnte Semantha so – deutlich schneller als Cem – die gesamten Vertragsunterlagen prüfen. Dabei werden auffällige Stellen markiert, z. B. wenn in der Vorschrift „gleichwertiges Erscheinungsbild“ und im Vertrag „moderne Ästhetik“ stehen. Diese Punkte kann Cem dann in Ruhe prüfen und entscheiden, ob die von der KI angemarkten Stellen tatsächlich eine Abweichung darstellen oder nicht. Fälle, bei denen das System sich unsicher ist oder Fälle, bei denen es eine klare Abweichung vermutet, können priorisiert bearbeitet werden.

In diesem Beispiel übernimmt die das textverarbeitende KI-System also einen Teil der menschlichen Aufgabe – die Sortierung der Stellen, die zu prüfen sind. Besonders erfahrene Sachbearbeitende können hier vielleicht auf Erfahrung zurückgreifen; ggf. gibt es auch interne Richtlinien, welche Stellen besonders geprüft werden müssen. Aber gerade in Fällen, wo die Masse an Text sehr groß ist, kann diese Priorisierung sehr schwierig sein und erfordert entweder viel Zeit oder muss zufällig vorgenommen werden.

Textverarbeitende KI wie Semantha kann die Verwaltung hierbei unterstützen, indem im Rahmen einer semantischen Analyse große Mengen an Daten untersucht und für die Bearbeitung durch menschliche Sachbearbeitende priorisiert werden.

10.6 Beispiele einfache Sprache

Im Rahmen dieses Fallbeispiels wollen wir uns noch einmal etwas detaillierter mit dem Konzept der „Einfachen Sprache“ im Verwaltungskontext und den möglichen Anwendungsbereichen von Algorithmen in Bezug auf diese beschäftigen.

Wie bereits zu den Zielen von Textverarbeitung erläutert, handelt es sich bei Einfacher Sprache um eine simplifizierte Variante unserer Alltagssprache. Sie wurde entwickelt, um eine möglichst bürgernahe und direkte Kommunikation zu ermöglichen und verzichtet dabei unter anderem auf komplexe Satzstrukturen und unnötige Fachbegriffe.

Die Zielgruppe der einfachen Sprache sind vor allem Menschen mit geringer sprachlicher Kompetenz, Personen mit geistigen Behinderungen oder Menschen, die Deutsch als Fremdsprache gelernt, aber noch kein hohes Sprachniveau erlangt haben. Vereinfacht gesagt geht es darum, die Behörden vom sperrigen Amtsdeutsch zu befreien und eine Kommunikationsform zu wählen, die möglichst zugänglich für alle Beteiligten ist. Einfache Sprache ist dabei vom Konzept der Leichten Sprache abzugrenzen, bei der es sich um eine nach festen Regeln agierende Sprachvariante handelt, die in den 1970ern explizit aus der Perspektive der Barrierefreiheit und Inklusion entwickelt worden ist (vgl. Kellermann, 2014).

Bei der Produktion und Konzeption von Texten in Einfacher Sprache kann der Einsatz von Algorithmen zur Textverarbeitung unterstützen. Wie zuvor bereits beschrieben wurde, können diese Algorithmen so trainiert werden, dass sie bereits existierende Textdateien zu neuen Texten umformen oder aber neue Texte aus Stichworten und Anmerkungen generieren. Dies ist genau der Anwendungsfall, der in Bezug auf Einfache Sprache benötigt wird. So können bereits bestehende Texte in komplexeren Sprachformen wie z. B. Amtsdeutsch in den mit Einfacher Sprache angelerten Algorithmus eingegeben werden, sodass dieser aus ihnen Texte mit den gleichen Informationen, aber einfacheren Satzstrukturen oder ähnlichen Merkmalen der Einfachen Sprache bildet. Dies ermöglicht es, bereits bestehende Informationen, die schon auf den Webseiten der Behörden existieren, anzupassen, ohne dass jeder einzelne Text händisch neu geschrieben werden muss, was einen großen Kosten- und Arbeitsaufwand mit sich bringen würde.

Ein weiterer Anwendungsfall, bei dem KI-Systeme unterstützen können, ist die Neuerstellung von Inhalten. Hier können Sprachtools wie Capito eingesetzt werden, die schon während des Schreibprozesses Informationen über den Komplexitätsgrad des Textes geben und es den Autor/innen so ermöglichen, sich an ihre Zielgruppen anzupassen. Die Einschätzung der Sprachkomplexität ist von

großer Bedeutung, wenn es beispielsweise darum geht, Texte mit wichtigen Informationen, wie das richtige Verhalten in Notfällen, für alle Adressaten zugänglich zu machen.

Grundsätzlich kann Einfache Sprache an allen Orten eingesetzt werden, an denen die Kommunikation mit einer sehr divers sprachkompetenten Gruppe von Personen gewährleistet werden muss. Je wichtiger die zu kommunizierenden Informationen sind, umso mehr sollte Einfache Sprache einbezogen werden. Es ist daher wenig überraschend, dass die Haupteinsatzgebiete vor allem auf den Webseiten und Informationsbroschüren von Behörden und Ämtern sowie medizinischen Einrichtungen zu finden sind.

Probleme von Einfacher Sprache bzw. deren technischer Einbindung durch Textverarbeitung sind vor allem die uneinheitlichen Regeln. Anders als bei Leichter Sprache gibt es keine zentral bestimmten Vorgaben für die Schreibweise. Vielmehr haben verschiedene Institutionen Richtlinien, die für Einfache Sprache genutzt werden können. Ein damit eng verbundener Problempunkt ist die zuvor vorgestellte Güte der Daten. Die Faktoren, nach denen ein lernender Algorithmus schlussendlich beurteilt, sind abhängig von den Datensätzen, mit denen er angelernt wird. Fehlerhafte oder von den gewünschten Regeln abweichende Datensätze könnten also zu unerwünschten bzw. falschen Ergebnissen führen.

Die Algorithmen, die für diese Art von Textverarbeitung angelernt werden, bedienen sich dabei der bereits in Abschn. „10.2 Grundlagen der Textverarbeitung“ vorgestellten Techniken zur Textverarbeitung. Grundlage bietet dabei vor allem die Entwicklung von Algorithmen auf Basis bestimmter Regelstrukturen. Beispielsweise kann ein Algorithmus, der lediglich aus der Anzahl der Worte pro Satz und der Länge eben dieser Worte einen Kennwert bildet, genutzt werden, um Aussagen über die Verständlichkeit des zu untersuchenden Textes zu treffen.

Eine weitere Möglichkeit ist der Einsatz von Deep Learning. Hier werden dem Algorithmus Trainingsdaten in großen Mengen zur Verfügung gestellt, auf deren Grundlage er trainiert wird und lernt, bestimmte Muster zu erkennen und Entscheidungen zu treffen. So könnte man bspw. großen Mengen an Texten in Einfacher Sprache zur Verfügung stellen, um dann schlussendlich neue Texte nach dem Konzept der Einfachheit bewerten zu lassen und so Stufen der Sprachkomplexität zu entwickeln. Diese Komplexitätseinschätzung durch den Algorithmus hilft dann den Nutzer/innen dabei, ihre neuen Textstücke an ihre Zielgruppen anzupassen. Auch die bereits vorgestellte Idee eines Übersetzungstools, bei dem normale Sprache in ein Feld eingegeben und Einfache Sprache ausgegeben wird, ließe sich mit dieser Art Algorithmen realisieren.

10.7 Aufgaben zum eigenen Anwendungsfall

In vielen Anwendungsfällen innerhalb der Verwaltung wird die Nutzung von textverarbeitenden Systemen eine wichtige Rolle spielen. Unterschiedliche Technologien sind dabei für bestimmte Aufgaben gerüstet – für andere nicht.

- Erörtern Sie, welche Schritte notwendig und zu prüfen sind, bevor Sie den Einsatz von textverarbeitender KI in Angriff nehmen. Bewerten Sie dabei einerseits den aktuellen Zustand aber auch den Arbeitsaufwand sowie etwaige Erleichterungen zur Integration einer textverarbeitenden KI in Ihrem Projekt.
- Diskutieren Sie, welche der vorgestellten Technologien eingesetzt werden können. Wählen Sie mindestens zwei Technologien aus und stellen Sie die Unterschiede anhand eines Anwendungsfalles dar. Wählen Sie dabei ein Beispiel aus Ihrem Projekt oder gehen Sie davon aus, dass mittels eines Chatbots Fragen von externen Partnern oder Bürger/innen beantwortet werden können sollen.
- Erstellen Sie für Ihren Anwendungsfall drei Beispielfragen und schildern Sie, was Intent, Entity & Kontext in diesen Fällen sein könnten.

Literatur

- Bürgernahe Verwaltungssprache. (2002). https://www.bva.bund.de/SharedDocs/Downloads/DE/Oeffentlichkeitsarbeit/Buergernahe_Verwaltungssprache_BBB.pdf?__blob=publicationFile&v=5. Zugegriffen: 15. Okt. 2022.
- Chilakapati, A. (23. September 2019). BoW to BERT. Data exploration. <https://xplordat.com/2019/09/23/bow-to-bert/>. Zugegriffen: 15. Okt. 2022.
- Jain, M., Kota, R., Kumar, P., & Patel, S. N. (2018). Convey: Exploring the use of a context view for chatbots. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (S. 1–6). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3174042>.
- Kellermann, G. (2014). Leichte und Einfache Sprache – Versuch einer Definition | APuZ. bpb.de, <https://www.bpb.de/apuz/179341/leichte-und-einfache-sprache-versuch-einer-definition>. Zugegriffen: 15. Okt. 2022.
- Russell, S., & Norvig, P. (2012). *Künstliche Intelligenz* (Bd. 2). Pearson Studium.

Weiterführende Literatur und Online-Artikel

- Alammar, J. (2018). The illustrated transformer, von <http://jalammar.github.io/illustrated-transformer/>. Zugegriffen: 15. Okt. 2022.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. <https://arxiv.org/abs/1810.04805v2>. Zugegriffen: 15. Okt. 2022.
- Roth, U. (2020). Künstliche Intelligenz schreibt in Einfacher Sprache/Plain Language, <https://leichtgesagt.eu/kuenstliche-intelligenz-schreibt-in-einfacher-sprache-plain-language>. Zugegriffen: 15. Okt. 2022.
- McTear, M. F., Callejas, Z., & Griol, D. (2016). The conversational interface (Bd. 6, No. 94, S. 102). Springer.
- Quamar, A., Lei, C., Miller, D., Ozcan, F., Kreulen, J., Moore, R. J., & Efthymiou, V. (2020). An ontology-based conversation system for knowledge bases. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 361–376. <https://doi.org/10.1145/3318464.3386139>.
- Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International journal of machine learning and cybernetics*, 1, 43–52.

Open Access Dieses Kapitel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Kapitel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

