Natural Language Processing:

Betrachtung automatischer Textzusammenfassung und damit verbundener Herausforderungen

Yoan Dimitrov
Fakultät Informatik, Studiengang Human-Centered Computing
Hochschule Reutlingen – Reutlingen University of Applied Sciences
Alteburgstr. 150, 72762 Reutlingen
yoan.dimitrov@student.reutlingen-university.de

ABSTRACT

Natural Language Processing beschäftigt sich mit der Interaktion zwischen Computern und natürlichen Sprachen, wie etwa Deutsch oder Englisch. Hierfür werden bewährte Methoden aus der Linguistik mit Techniken aus der Informatik und künstlicher Intelligenz fusioniert, um eine möglichst intelligente Kommunikation zu erreichen. Diese Hausarbeit hat das Ziel, einen Überblick über die Grundlagen, Anwendungsgebiete und Herausforderungen von *Natural Language Processing* zu vermitteln. Darauf bauend wird das Anwendungsgebiet – automatische Textzusammenfassung – vorgestellt und insbesondere hinsichtlich dessen Funktionsweise näher untersucht.

CCS CONCEPTS

• Artificial intelligence → Natural language processing

KEYWORDS

Natural Language Process, Machine learning, Automatic Text Summarization, Summarization Generation

1 Einleitung

Natural Language Processing (deutsch: natürliche Sprachverarbeitung, kurz: NLP) ist ein umfangreiches Forschungsgebiet der Informatik, welches sich seit den 1950er Jahren auf die computerbasierte Erfassung und Weiterverarbeitung natürlicher Sprachen fokussiert. Diese Disziplin beschränkt sich jedoch nicht nur auf Informatik, sondern umfasst unter anderem Konzepte und Techniken aus der Linguistik, kognitiven Psychologie und künstlichen Intelligenz [7, 11].

Ein NLP-System zielt darauf ab, natürliche Sprachen computergestützt zu modellieren [7]. Ohne solche NLP-Systeme wären Anwendungen, wie etwa Suchmaschinen, Dialogsysteme[1], automatische Übersetzungen[2] oder Textzusammenfassungssysteme nicht realisierbar.

Als natürliche Sprache wird jede Sprache bezeichnet, welche Menschen verwenden, um miteinander zu kommunizieren. Mittels solcher Sprachen drücken Menschen einerseits ihre Gefühle und Wissen aus. Andererseits vermitteln sie anderen Menschen ihre Gedanken und Antworten [7, 11, 14].

NLP nimmt zwei zentrale Aufgaben in Angriff. Erste Aufgabe des NLP ist das Verstehen der Bedeutung einer vom Computer erfassten natürlichen Sprache. Hierbei wird eine natürliche Sprache häufig zunächst mittels Spracherkennung in Text umgewandelt. Anschließend wird versucht, die Semantik des Texts herauszufinden [7, 11]. Dieser Prozess wird in der Literatur mit dem Begriff *Natural Language Understanding* bezeichnet [11]. Das zweite Aufgabenfeld beschäftigt sich mit der Fragestellung, wie Computer Sprachen selbst erzeugen können. Dies wird auch *Natural Language Generation* genannt [11]. Grundlage hierfür bilden die Ermittlung, Musteridentifizierung und Formatierung von Daten aus unterschiedlichen Quellen. Auf dieser Basis lassen sich später menschenähnliche Sätze in Form von Text oder Sprache erzeugen [7, 11]. In der Praxis erfüllen NLP-Systeme je nach Problemstellung eine oder beide der vorgenannten Aufgaben. Beide Aufgaben setzen jedoch Verarbeitungsverfahren aus der Informatik, Mathematik und Sprachwissenschaften voraus, um durchgeführt werden zu können.

In folgender Arbeit werden die elementaren Charakteristika, Verarbeitungsverfahren, Anwendungsgebiete und Herausforderungen des NLP vorgestellt. Da es im Rahmen dieser Arbeit unmöglich ist, das gesamte Spektrum von theoretischen und praktischen Techniken des NLP zu adressieren [7], wird nur ein Einblick von der Vielfalt der in den letzten 60 Jahren entwickelten Verarbeitungsverfahren vermittelt. Nach einer Einführung in NLP, wird der Fokus dieser Arbeit auf die signifikanten Aspekte und

[1] https://www.apple.com/de/siri/ Abruf: 22.01.19

[2] https://translate.google.com Abruf: 22.01.19

Schwierigkeiten des Anwendungsgebietes *Automatische Textzusammenfassung* gelenkt.

1.1 Aufbau der Ausarbeitung

In Kapitel 2 erfolgt eine Einführung in die Grundlagen von NLP. Zunächst werden linguistische Ebenen aus der Sprachwissenschaft im Kontext von NLP vorgestellt. Anschließend folgt ein Einblick in die gängigen Verarbeitungsverfahren von NLP. Darauf aufbauend werden Anwendungsgebiete vorgestellt und kurz beschrieben. Das Kapitel endet mit einem Überblick über die Herausforderungen von NLP. Kapitel 3 setzt sich mit dem Anwendungsgebiet *Automatische Textzusammenfassung* auseinander. Nach einer kurzen Einführung werden verschiedene Verfahren zur automatischen Textzusammenfassung analysiert und hinsichtlich deren Eigenschaften und Herangehensweise verglichen. Darauffolgend wird auf die Anwendungen, Einschränkungen und Herausforderungen eingegangen. Abschließend sollen in Kapitel 4 die Ergebnisse dieser Arbeit zusammenfassend dargestellt werden.

2 Grundlagen

2.1 Linguistische Ebenen

Zur effizienten Interpretation einer natürlichen Sprache lösen NLP-Systeme eine Reihe an Probleme. Jedes Problem stellt dabei eine Ebene der Linguistik dar. Dazu gehören: die Phonologie-Ebene, Morphologie-Ebene, lexikalische Ebene, Syntaktik-Ebene, Semantik-Ebene, Diskurs-Ebene und die pragmatische-Ebene [11, 17]. Da Menschen alle Sprachebenen verwenden, um ein optimales Verständnis über einen Satz zu erlangen, soll ein NLP-System ebenfalls möglichst alle Sprachebenen beherrschen können [11]. Nachfolgend werden die einzelnen Ebenen kompakt erläutert.

Die Phonologie-Ebene behandelt die Interpretation von Sprechlauten innerhalb und quer durch Wörter. Dabei werden Schallwellen untersucht und in ein digitales Signal umgewandelt. Darauf basierend können beispielsweise wichtige Hinweise über die Bedeutung eines Wortes interpretiert werden. NLP-Systeme wenden diese Ebene an, wenn der Ursprung einer Eingabe einer gesprochenen Sprache entspricht [11].

Die Morphologie- und lexikalische Ebenen sind eng miteinander verwandt. Sie analysieren die Zusammensetzung von Wörtern und deren untereinander bestehenden Beziehungen. NLP-Systeme benötigen diese Ebenen, um die Merkmale eines Wortes zu bestimmen. Dazu zählen die Präfixe, Suffixe, individuelle Bedeutung und Wortart (*part of speech*, POS) eines Wortes. Zu erkennen gilt beispielsweise, ob ein Wort ein Verb oder Adjektiv darstellt. Je nach Position in einem Satz kann ein Wort verschiedene Funktionen einnehmen [11, 17].
Auf der Syntaktik-Ebene analysieren NLP-Systeme mehrere Wörter in einem Satz, um die Struktur des Satzes hinsichtlich der Grammatik zu verstehen. Ziel ist es sicherzustellen, dass jeder Satz den vordefinierten Grammatikregeln entspricht. Dabei spielen die Reihenfolge der Wörter und deren Relationen zueinander eine entscheidende Rolle [11, 17].

Semantische Verarbeitung wird eingesetzt, um die möglichen Bedeutungen eines Satzes zu bestimmen. Der Fokus liegt dabei auf der Interaktion von Wort-Sequenzen [11, 17]. Eine besondere Herausforderung an dieser Stelle stellt die Mehrdeutigkeit von Sätzen dar [17].

Auf der Diskurs-Ebene stellen sich NLP-Systeme die Frage, wie die einzelnen Sätze eines Texts zusammenhängen. Hier existieren mehrere Diskurstheorien. Ihr Ausgangspunkt liegt in der Idee, dass ein Text Beziehungen zwischen Sätzen umfasst, die seine Kohärenz ergänzen oder sogar bestimmen [11, 17].

Die pragmatische Ebene befasst sich mit der Sprachverwendung in einem Kontext. NLP-Systeme untersuchen dabei die Äußerungen eines Textes, um die impliziten Absichten, Ziele und Pläne des Sprechers aufzudecken. Voraussetzung hierfür ist unter anderem möglichst viel Wissen über allgemeinen Sprachgebrauch und Konventionen [8, 11, 17].

Es ist anzumerken, dass NLP-Systeme nicht unbedingt alle linguistischen Ebenen berücksichtigen müssen. Diese Entscheidung hängt von der jeweiligen Anwendung und deren Zielen ab. Folgende Abbildung stellt die Sprachebenen und deren Anwendung übersichtlich dar.

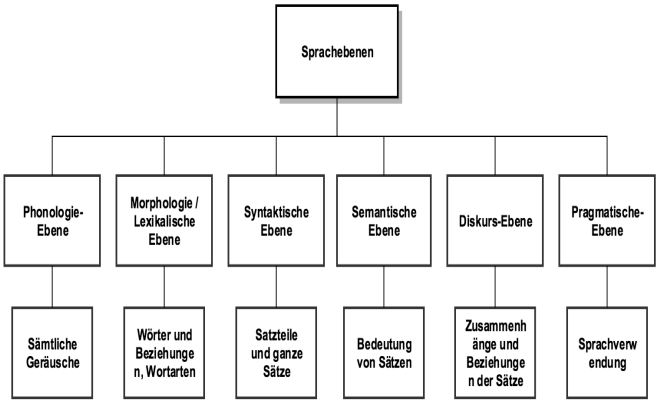

Abbildung 1: Sprachebenen und deren Aufgaben[3] [8, 11, 17]

[3] Vergleich Schaubild: https://www.uni-due.de/SHE/REV_Levels_Chart.htm Abruf: 22.01.19

2.2 Verarbeitungsverfahren

Die verschiedenen Ebenen einer Sprache, die im vorherigen Abschnitt beschrieben wurden, werden durch die Verwendung diverser Verarbeitungsverfahren beziehungsweise formaler Modelle erfasst [8]. Im Bereich des NLP ist allgemein zwischen drei Arten von Verarbeitungsverfahren zu unterscheiden: regelbasierten Verfahren, statistischen Verfahren und neuronalen Netzen. In der Praxis werden jedoch häufig Hybrid-Ansätze entwickelt, welche gleichsam die Stärken der drei Verfahren bündeln [11]. Nachfolgend werden die einzelnen Verfahren vorgestellt und diskutiert.

2.2.1 Regelbasierte Ansätze

Die frühesten Verarbeitungsverfahren basierten auf Regeln, welche meistens von Domänenexperten formuliert wurden [7, 8, 11, 14, 21]. Als ausgereifte Beispiele regelbasierter Ansätze sind reguläre Ausdrücke und kontextfreie Grammatiken zu nennen [7, 8, 14]. Ein Anwendungsbeispiel stellt die Übersetzung der Sprache Englisch in Deutsch dar[4]. Dazu sind nämlich Grammatikregeln notwendig, um die englische und deutsche Satzstruktur abzubilden. Hier ist denkbar, Regeln zu definieren, welche die Funktion eines Wortes, zum Beispiel Nomen oder Verb, erkennen. Eine Zuordnung zwischen englischen und deutschen Worten ist ebenfalls erforderlich. Hierzu wird eine Art Lexikon herangezogen. Schließlich müssen Regeln definiert werden, welche die Satzstrukturen beider Sprachen in Beziehung setzen können. Herausforderungen regelbasierter Ansätze tauchen beispielsweise bei der Extrahierung der Semantik eines Textes auf, da formale Grammatiken nur eine gültige Syntax spezifizieren. In der Praxis entsteht jedoch üblicherweise die Problematik der Mehrdeutigkeit eines Wortes. Daher spielt die Semantik eines oder mehrerer Wörter eine wesentliche Rolle und muss berücksichtigt werden. Eine weitere Herausforderung liegt in der Frage, ob sich Regeln, welche meist kontextspezifisch definiert werden, ohne erheblichen Aufwand generalisieren lassen, um eine Wiederverwendbarkeit zu ermöglichen. Nachteilig erweist sich, dass enorme Zeit benötigt wird, um alle Regeln eines NLP-Systems zu schreiben und verwalten. Desweitern sind regelbasierte Ansätze auf das Fachwissen der Domänenexperten stark angewiesen. Ein Vorteil gegenüber anderen Verfahren lässt sich jedoch an der Genauigkeit regelbasierter Ansätze erkennen. Regel werden für spezifische Szenarien formuliert und erfüllen ihren Zweck aufgrund dieser Eingrenzung sehr präzise [7, 11, 14].

2.2.2 Statistische Ansätze

Statistische Ansätze adressieren die Herausforderungen der maschinellen Verarbeitung natürlicher Sprachen anhand verschiedener mathematischer Techniken [8, 11, 14, 21]. Diese Ansätze werden häufig präferiert, wenn zu viele Fälle und Ausnahmen existieren, die mit handgeschriebenen Regeln schwierig zu behandeln sind [11, 14]. Wahrscheinlichkeiten und große, Datensätze bilden deren Grundlage. NLP-Systeme, welche statistische Verfahren als Basis der Sprachverarbeitung anwenden, stellen die Frage in den Vordergrund, was wahrscheinliche und unwahrscheinliche Strukturen eines Textes sind. Dafür werden von menschlichen Experten ausgesuchte Merkmale[5] aus Datensätzen analysiert, um Wissen aufzubauen und eigene Regeln daraus abzuleiten. Hier ist jedoch vorausgesetzt, dass große, strukturierte Datensätze bereits vorliegen. Daher stellen üblicherweise die Datenbeschaffenheit und -vorbereitung die größten Hürden bei dieser Kategorie dar. Statistische Ansätze bringen aber auch unmittelbare Vorteile hinsichtlich deren Lernfähigkeit mit sich: das manuelle Schreiben von Regeln und Grammatiken ist nicht erforderlich. [8, 11, 14, 21].

Generell existiert eine Vielfalt an statistischen Methoden, die unterschiedliche Aufgaben entsprechend den linguistischen Ebenen erfüllen [8, 14, 21]. Eine kleine Auswahl wird im Folgenden kurz vorgestellt. Zur weiteren Vertiefung der verschiedenen, statistischen Verfahren sei auf Grundlagenbücher, zum Beispiel von [8], verwiesen.

N-Gramm: ist, vereinfacht gesagt, eine Folge von N-Wörtern, wobei N die Anzahl der Wörter repräsentiert. Der Ausdruck „Mein Name“ wird beispielsweise als ein 2-Gramm bezeichnet. Folglich ist „Er geht nach“ logischerweise ein 3-Gramm. Grundsätzlich versucht ein N-Gramm-Modell das nächste Wort basierend auf dem Auftreten seiner N - 1 vorhergehenden Wörter vorherzusagen. Zum Beispiel kann ein 3-Gramm-Modell das nächste Wort auf Basis seiner vorherigen zwei Wörter vorhersagen, da in diesem Fall N-1= 2 gilt [8, 14].

Hidden-Markov-Model (HMM): ist ein probabilistisches Sequenzmodell. Basierend auf früheren Entscheidungen, wie zuvor erkannten Wörtern, Buchstaben oder Sätzen, und einem aktuellen Datensatz versucht ein HMM die wahrscheinlichste Entscheidung zu treffen. Häufig wird HMM eingesetzt, um die Wortarten von einer Sequenz von Wörtern richtig zu identifizieren [8, 14]. HMM wird außerdem im Bereich der Spracherkennung verwendet. Ziel dabei ist die wahrscheinlichste Folge von Phonemen anhand der Wellenform eines gesprochenen Wortes herauszufinden [14].

Naive Bayes: ist ein Wahrscheinlichkeitsalgorithmus, der den Satz von Bayes verwendet, um die Klasse eines Textes vorherzusagen. Naive Bayes trifft die Annahme, dass die einzelnen Merkmale eines Datensatzes voneinander unabhängig sind. Das Text-Dokument wird als ein *Bag-of-Words* dargestellt. Das bedeutet, dass die Reihenfolge von

[4] https://datascience.stackexchange.com/questions/18100 Abruf: 22.01.19

[5] Auch Feature-Engineering genannt

Wörtern ignoriert wird. Die Häufigkeit der Wörter bleibt im Dokument jedoch erhalten [8]. Mit diesem Algorithmus können beispielsweise Online-Artikel klassifiziert werden, indem die Wahrscheinlichkeit berechnet wird, dass die Klasse des Artikels entweder Sport oder Politik ist.

2.2.3 Neuronale Netze

Neben den klassischen, aber nicht weniger anspruchsvollen, Verarbeitungsverfahren, die bereits thematisiert wurden, haben sich in letzten 30 Jahren auch neuronale Netze als potentielle Alternative erwiesen. Derzeit herrscht ein Hype über neuronale Netze und die damit verbundenen Möglichkeiten. Diese Popularität ist auf zwei wesentliche Aspekte zurückzuführen: die riesigen Datenmengen, welche heute zur Verfügung stehen und die verbesserte Rechenleistung[6]. Im Wesentlichen versuchen neuronale Netze anhand miteinander verbundener künstlicher Neuronen die Funktionsweise des Gehirns zu imitieren, um somit Zusammenhänge und Merkmale aus natürlichen Sprachen ebenfalls erfassen und verarbeiten zu können. Sie enthalten jedoch keine expliziten Regeln [6, 8, 11, 21, 22]. Analog zu statistischen Ansätzen können neuronale Netze aus Daten lernen. Die einfachste Art von neuronalen Netzen ist das *Feed-Forward-Netzwerk* [8, 21, 22] und wird in Abbildung 2 veranschaulicht.

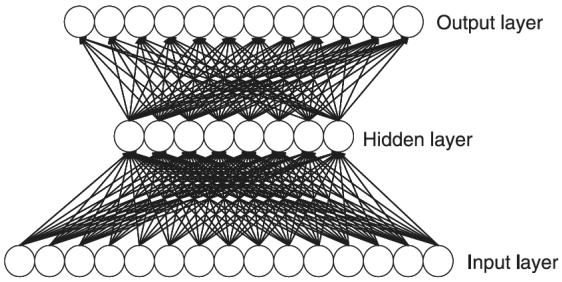

Abbildung 2: Feed-Forward-Netzwerk [22]

Ausgehend vom Input-Layer werden Daten, beispielsweise Wörter in Form von Vektoren, anhand sogenannter Gewichte ausgewertet und gegebenenfalls modifiziert, bis sie als Ausgabe klassifiziert werden können. Die Ausgänge der Neuronen jeder Schicht werden an Neuronen in der nächsthöheren Schicht übergeben. Beim Feed-Forward-Netzwerk werden keine Ausgaben den unteren Schichten zurückgegeben [8, 21, 22]. Da ein umfassender Vergleich neuronaler Netze und deren Funktionsweise nicht Kern dieser Ausarbeitung ist, wird an dieser Stelle auf eine weitergehende Recherche verzichtet. Für einen tieferen Einblick sei auf *Speech and Language Processing* [8] verwiesen.

Ein entscheidender Vorteil von neuronalen Netzen gegenüber anderen Ansätzen ergebt sich in der Fähigkeit Klassifizierungs- und Regressionsaufgaben mit Datensätzen, welche nichtlineare oder komplexere Beziehungen aufweisen, effizient durchzuführen. Die Abwesenheit vom aufwendigen Feature-Engineering stellt einen weiteren Vorteil dar. Ein Nachteil zeigt sich darin, dass es nur mit erheblichem Aufwand nachvollziehbar ist, weshalb ein neuronales Netz ein bestimmtes Ergebnis ausgibt. Ähnlich wie statistische Ansätze, stellen die Datenbeschaffenheit sowie die Referenzdaten-gewinnung bei neuronalen Netzen eine weitere, zu überwindende Herausforderung dar [6, 8, 20, 22].

2.3 Anwendungsgebiete

Das Einsatzspektrum von NLP ist breit aufgestellt und umfasst diverse Anwendungsgebiete. Einige Anwendungsbeispiele werden an dieser Stelle kurz vorgestellt und beschrieben.

Question-Answering sind Systeme, welche die von einem Benutzer gestellten Fragen in einer natürlichen Sprache beantworten. Im engen Zusammenhang hierzu steht *Information Retrieval*, eine Technik zur Rückgewinnung von für einen Benutzer relevanten Informationen [8, 11].

Dialogue Systems, wie bereits zuvor erwähnt, interagieren mit einem Benutzer in einer natürlichen Sprache in Form von Text, Sprache oder einer Kombination aus beiden [8, 11]. Hier wird zwischen *Task-Orientated Dialog Agents*, welche sich für bestimmte Aufgaben und kürzere Gespräche eignen und *Chatbots*, welche die Unterhaltung mit Menschen oder anderen technischen Systemen erlaubt, unterschieden [8].

Machine Translation stellt eine alte Technik des NLP dar und dient der Übersetzung eines Textes von einer Sprache in eine andere [11]. Ein bekanntes, ausgereiftes Beispiel hierfür ist *Google Translate*[7].

Bei *Spam Detection* handelt sich um die binäre Klassifizierung einer E-Mail. Typischerweise wird eine E-Mail der Klasse Spam oder Nicht-Spam zugeordnet [8].

Automatic Text Summarization hat die Aufgabe, eine übersichtliche, strukturierte Zusammenfassung eines Texts zu erstellen, während der Inhalt und die Semantik der wichtigsten Informationen erhalten bleiben [1, 9, 11, 16]. Im weiteren Verlauf der Ausarbeitung wird dieses Themengebiet näher untersucht.

2.4 Herausforderungen

NLP-Systeme werden mit einer Vielzahl verschiedener Herausforderungen konfrontiert. Besonders hervorzuheben ist

[6] https://towardsdatascience.com/hype-disadvantages-of-neural-networks-6af04904ba5b Abruf: 22.01.19

[7] https://translate.google.com/ Abruf: 22.01.19

die korrekte Erkennung von Ironie und Sarkasmus sowie anderen rhetorischen Stillmitteln [5]. Diese Problemstellungen haben bereits tiefgehende Studien motiviert. So haben González-Ibáñez et. al. grundlegende Klassifizierungsmethoden, wie *Support-Vector-Machines* und *Logistic Regression*, untersucht mit dem Ziel, Sarkasmus in positiven beziehungsweise negativen Twitter-Nachrichten automatisch erkennen zu können. Dabei wurden jedoch nur lexikalische und pragmatische Faktoren berücksichtigt. Dies hat dazu geführt, dass eine präzise Unterscheidung von Sarkasmus in positiven und negativen Twitter-Nachrichten nicht möglich war. Festgestellt wurde, dass weitere Informationen über die Interaktion zwischen Twitter-Nutzern erforderlich sind, um die korrekte Erkennung von Sarkasmus zu ermöglichen [5].

Wie bereits zuvor erwähnt, stellt die Mehrdeutigkeit von Worten und Sätzen eine weitere Herausforderung dar. Ausnahmen hierzu existieren nicht; alle linguistischen Ebenen sind betroffen. Beispielsweise sind Wörter je nach natürlicher Sprache lexikalisch mehrdeutig. Dies heißt konkret, dass ein Wort mehreren Wortarten zugeordnet werden kann. Auch ganze Sätze können eine Mehrdeutigkeit in vieler Hinsicht aufweisen. Darüber hinaus unterscheidet sich häufig die Bedeutung von gleich ausgesprochenen Worten oder Wortfolgen[8] [7, 8, 11, 14].

Aufgrund der zunehmender Mensch-Computer-Interkation in dieser Disziplin sind ethische Problemfelder ebenso zu behandeln. [9]. Leidner et. al. adressieren unter anderem NLP-Applikationen, welche in der Vergangenheit gegen ethische Grundsätze verstoßen haben. Auf Twitter wurden beispielsweise *Chatbots* bereits eingesetzt, um politische Propaganda zu betreiben. Außerdem wurde offengelegt, dass viele NLP-Applikation private, persönliche Attribute und Eigenschaften aus Benutzerdaten ermitteln, welche in Wirklichkeit dem Datenschutz unterliegen [9]. Aspekte wie Fairness und Diskriminierung sind ebenfalls in Bezug auf ethische NLP-Applikationen Gegenstand aktueller Forschung. Es stellt sich beispielsweise die Frage, wie Dialogsysteme mit Menschen umgehen sollen, welche mit einem starken Akzent oder schwierigen Dialekt sprechen. Solche Angelegenheiten müssen weiter untersucht werden, um deren Auswirkungen auf die Gesellschaft herauszufinden [9].

3 Automatische Textzusammenfassung

Das Kapitel widmet sich der zweiten Zielsetzung dieser Ausarbeitung: die Betrachtung automatischer Textzusammenfassung und damit verbundener Herausforderungen. Zunächst wird auf die Grundlagen und Verarbeitungsverfahren von automatischer Textzusammenfassung eingegangen. Danach erfolgt eine Analyse bestehender Anwendungsdomänen und möglicher Herausforderungen.

3.1 Grundlagen

Allgemein dienen Zusammenfassungssysteme der Reduzierung eines Textes auf den wesentlichen Inhalt, der dem Leser die Schlüsselinformationen und Bedeutung in kompakter Form übermittelt [1, 16, 19]. Die Textzusammenfassung stellt jedoch eine nicht triviale Aufgabe dar. Während Menschen lesen, verstehen und auf frühere Erfahrungen zurückgreifen können, um einen Text zusammenzufassen, fehlt dem Computer häufig das menschliche Wissen und sprachliche Erfahrung [1]. In den letzten Jahren wurden beträchtliche Fortschritte erzielt, um den Herausforderungen der automatischer Textzusammenfassung zu begegnen.

In aktueller Literatur werden Textzusammenfassungssysteme häufig anhand mehrerer Kriterien kategorisiert [1, 9, 16, 19]:

Als wichtiges Kriterium ist der Eingabetyp einzustufen. Dabei wird zwischen einem oder mehreren zusammenzufassenden Texten unterschieden. Letztere werden zu einem einzigen Dokument gefasst. Dabei wird häufig davon ausgegangen, dass sämtliche Texte, auch Multi-Dokumente genannt, die gleichen oder ähnlichen Themen diskutieren. Insbesondere im Web-Kontext sind die Multi-Dokument-Zusammenfassungen sehr gefragt, um dem Leser einen kompakten Überblick über bereits verfasste Texte eines bestimmten Themas zu geben [1, 16].

Ein Zusammenfassungssystem verfolgt in der Regel einen domänenspezifischen, query-basierenden oder generischen Verwendungszweck. Für den domänenspezifischen Zweck ist häufig eine Wissensdatenbank zur Unterstützung der Satzauswahl beim Zusammenfassungsprozess erforderlich. Fachartikel über Politik, Medizin und Finanzen oder Wetterberichte stellen typische, domänenspezifische Anwendungsfälle dar. Query-basierte Zusammenfassungen enthalten hingegen ausschließlich Informationen, welche von einem Benutzer in Form von Fragen oder Stichworten abgefragt wurden. Diese beziehen sich immer auf ein bestimmtes Thema. Bei einer generischen Zusammenfassung wird ein Text unabhängig von seinem Thema oder seiner Domäne zusammengefasst. Im konkreten Fall werden keine Annahmen über die Domäne der Quelleninformationen getroffen. Alle Dokumente werden als homogene Texte betrachtet [16, 19].

Ein weiteres Kriterium zur Unterscheidung von Zusammenfassungssystemen ist deren Ausgabetyp. Einerseits besteht die Möglichkeit, wichtiger Sätze in einem Text zu identifizieren und diese als Zusammenfassung auszugeben. Sämtliche Sätze werden aus dem Text direkt extrahiert. Dieser Ansatz wird als *Extract Summarization* bezeichnet und hat in der Vergangenheit relativ gute Ergebnisse erhalten. Dem gegenüber existieren abstrakte Zusammenfassungsmethoden

[8] Zum Beispiel „wahr“ und „war“

(englisch: Abstract Summarization). Sie zielen darauf ab, einen Text mit geeigneten Verarbeitungsverfahren zu untersuchen und daraufhin einen neuen, kürzeren Text zu erzeugen, welcher die wichtigsten Informationen des Originals enthält. Von Menschen erstellten Zusammenfassungen fallen üblicherweise unter diese Kategorie. Aus technischer Sicht sind abstrakte Zusammenfassungssysteme ohnehin schwierig zu realisieren, da mehrere Aufgaben, wie zum Beispiel Sprachgenerierung oder semantische Verarbeitung, abgearbeitet werden müssen. Hierbei gilt es auch zu erwähnen, dass sich der Ausgabetyp der Zusammenfassung unmittelbar auch auf die zugrunde liegenden Verarbeitungsverfahren auswirken [1, 9, 16, 19].

3.2 Verarbeitungsverfahren

Nachfolgend werden einige Verarbeitungsverfahren und Techniken für beide der zuvor eingeführten Ausgabetypen vorgestellt.

Extract Summarization

Der Fokus vieler wissenschaftlichen Arbeiten liegt auf den Verarbeitungsverfahren zur Realisierung von *Extraction Summarizer*s [1, 3, 4, 9, 13, 16, 19].

Ziel ist, wie zuvor erwähnt, wichtige Sätze aus einem Text zu extrahieren, um daraus eine Zusammenfassung zu erzeugen [9]. Hierfür werden statistische Verfahren eingesetzt, welche über die Jahre gereift sind. Edmundson bildete bereits 1968 die Grundlage für den Einsatz von statistischen Verfahren. Er definierte drei Merkmale, welche als Indikator für die Relevanz der Sätze gelten. Dazu gehören die Satzposition, ein vorhandenes Titelwort und Stichwörter [3, 16]. Die Sätze, welche zur Erstellung der Zusammenfassung dienen, werden anhand der identifizierten Merkmale bewertet. Hierfür erhalten die Merkmale Gewichte, um ihre Wichtigkeit zu bestimmen [3].

Weitere Verarbeitungsverfahren setzen auf die Berechnung der Worthäufigkeit innerhalb eines Satzes, um eine Zusammenfassung zu erzeugen. Ein trivialer Lösungsweg ist die tatsächliche Berechnung der Worthäufigkeit, indem das Wortvorkommen einfach summiert wird. Aussagekräftig ist dieser Ansatz jedoch häufig nicht, da die Textlänge die Worthäufigkeit stark beeinflussen kann. Um diese Hürde zu überwinden, wird auf *term frequency-inverse document frequency* (tf-idf) zurückgegriffen. Dieses Verfahren stellt die Frage in den Vordergrund, ob alle Wörter, welche in einem Text vorkommen, gleich wichtig sind. Die Idee dahinter ist die Reduzierung des Gewichts häufig vorkommender Wörter. Um dies zu erreichen, wird die proportionale Häufigkeit solcher Wörter mit der Häufigkeit derselben Wörter aus anderen zusammenhängenden Texten verglichen [1, 9, 13, 16].

Erkan et. al. führten einen Graph-basierten Ansatz zur Textzusammenfassung ein [4]. Im Kontext von Textzusammenfassung, repräsentieren die Knoten eines Graphs Sätze und jede Kante entspricht dem Gewicht zwischen Sätzen [9]. Nach Erstellung des Graphs werden wichtige Sätze identifiziert. Sätze, welche mit vielen anderen Sätzen verbunden sind, werden für die Zusammenfassung als essentiell erachtet [4].

Neben obengenannten Verfahren haben sich *bayessche* Verfahren im Bereich des maschinellen Lernens als geeignete Lösung für *extract summarization* erwiesen. Basierend auf Satzmerkmalen wird die Wahrscheinlichkeit berechnet, dass ein Satz zur Zusammenfassung hinzugefügt werden soll [9, 13, 16].

Abstract Summarization

In den letzten Jahren wurden erhebliche Fortschritte speziell im Bereich von neuronalen Netzen erzielt, um den Herausforderungen der abstrakten Textzusammenfassung gerecht zu werden. Bei diesem Ansatz werden zum einen entweder neue Sätze generiert oder vorhandene Sätze umformuliert. Zum anderen werden Vokabeln eingesetzt, welche nicht im ursprünglichen Text enthalten sind [2, 8, 15, 18].

Chopra et. al. führten eine Architektur in Form eines *Recurrent Neural Network* (RNN) ein, um Schlagzeilen basierend auf der ersten Zeile eines Nachrichtenartikels zu generieren [2]. RNNs haben im Gegensatz zu herkömmlichen neuronalen Netzen die Fähigkeit, sich an Informationen zu erinnern, die sie zuvor berechnet haben. Sie eignen sich daher ideal, um ein Wort basierend auf seinen vorherigen Wörtern vorherzusagen [2, 8]. RNNs sind jedoch nicht nur auf abstrakte Zusammenfassungssysteme eingeschränkt, sondern kommen in diversen Bereichen der künstlichen Intelligenz immer öfter zum Einsatz, um Abfolgen vorherzusagen.

Unter Verwendung von *CNN*- und *DailyMail*-Daten haben Nallapati et. al. ebenfalls RNNs eingesetzt und trainiert. Ziel war, ausgehend von großen, inhaltsreichen Dokumenten, prägnante Zusammenfassungen mit mehreren Sätzen zu erzeugen. Auch hier wird ein Vokabular verwendet, welches in den ursprünglichen Dokumenten nicht vorhanden ist. Im Rahmen der Bewertung der vorgeschlagenen Modelle wurde eine deutliche Leistungssteigerung bei Multi-Dokument-Zusammenfassungen festgestellt [15].

In der Arbeit von Rush et. al. wird ein *Attention-Based Summarization*-Ansatz demonstriert, welcher sich auf die Zusammenfassung auf Satzebene konzentriert und eine Vorbereitung von einem speziellen Vokabular nicht vorsieht. Bestrebt wird, die Beziehungen zwischen Sätzen vollständig zu erfassen [18]. Traditionell wird vom neuronalen Netz ein

ganzer Satz eingelesen; daraufhin werden alle Informationen komprimiert und in einen Vektor fester Länge eingetragen. Dies hat in der Vergangenheit jedoch öfters zum Verlust wichtiger Informationen geführt [2, 18]. Rush et. al. haben daher ihren *Attention*-Mechanismus vorgeschlagen, um alle Informationen des Satzes zunächst einzusehen und entsprechend dem aktuellen Kontext genaue, abstrakte Zusammenfassungen zu erstellen [18].

3.3 Anwendungsdomänen

Die stetige Weiterentwicklung und Verbesserung von Textzusammenfassungssystemen resultieren in neuen Anwendungsfällen. Diese sind überwiegend auf bestimmte Domänen zugeschnitten. So konzentrieren sich beispielsweise Chopra et. al. und Nallapati et. al. auf Nachrichtenartikel beliebter Nachrichtensender [2, 15]. Andere Anwendungsgebiete adressieren die Zusammenfassung von Webseiten, wissenschaftlichen Arbeiten, Emails und Blogs [1, 9, 16]. Auch im medizinischen Bereich finden Zusammenfassungssysteme Anwendung. Somit können Ärzte beispielsweise eine kurze, prägnante Übersicht über die medizinische Vorgeschichte ihrer Patienten erhalten [1, 9]. Einige fortgeschrittene Zusammenfassungssysteme zur generischen Zusammenfassung, wie etwa *Agolo*[9] oder *Text Summarization API*[10] von *DeepAI*, sind bereits öffentlich verfügbar und weisen gute Ergebnisse vor. Losgelöst von einem Anwendungsgebiet lassen sich die Vorteile von Zusammenfassungssysteme klar erkennen.

3.4 Herausforderungen

Generell gelten die in Kapitel 2.4 aufgeführten Herausforderungen auch für Textzusammenfassungssysteme. Es existieren jedoch zusätzliche, kontextspezifische Herausforderungen. Lloret et. al. haben sich in ihrer Arbeit mit über die bekannten Herausforderungen hinausgehenden Problemstellungen befasst [12]. Ein typisches Problem liegt darin, dass sich zwei Sätze aus unterschiedlichen Wörtern zusammensetzen aber semantisch äquivalent sind. Hierdurch stellt sich eine erhebliche Schwierigkeit bezüglich der Erkennung relevanter Sätze. Es wird außerdem suggeriert, dass automatische Textzusammenfassungen anhand von Menschen geschriebenen Zusammenfassungen hinsichtlich deren informativen Qualität bewertet werden sollen. Hierbei ergeben sich mehrere Herausforderungen, denn der Inhalt einer vom Menschen geschriebenen Zusammenfassung hängt von subjektiven Aspekten, wie seiner Ausbildung oder seinem aktuellen Zustand, ab. Die Lesbarkeit**,** die Schreibweise und die Grammatik von Zusammenfassungen spielen dabei eine wesentliche Rolle. Trotz Fortschritte zur Automatisierung dieser Aspekte, ist es weiterhin nötig, dass ein Fachexperte zur Überprüfung der automatischen Textzusammenfassung herangezogen wird [12].

4 Zusammenfassung

Entsprechend der am Anfang definierten Zielsetzung wurde ein umfassender Überblick über *Natural Language Processing* (NLP), ein Untergebiet der künstlichen Intelligenz, gegeben. Das Gebiet ist in Natural *Language Understanding* und *Natural Language Generation* einzuteilen. NLP-Systeme lösen, wie in Kapitel 2 dargelegt wurde, eine Reihe an Aufgaben entsprechend den linguistischen Ebenen, wobei sich die Aufgabenstellungen je nach Kategorie unterscheiden können. Diverse Verarbeitungsverfahren werden dafür eingesetzt und lassen sich in drei Gruppen einstufen: regelbasierte Ansätze, statistische Ansätze und neuronale Netze. Für den praktischen Einsatz werden öfters Hybrid-Ansätze implementiert, um die Stärken der einzelnen Verfahren zu kombinieren. Traditionelle Anwendungsgebiete von NLP umfassen beispielsweise *Dialogue Systems, Question-Answering Systems* oder *Automatic Text Summarization Systems*. Es existieren viele mit NLP-Systemen zusammenhängende Herausforderungen. Hierzu gehören die Mehrdeutigkeit natürlicher Sprachen oder die korrekte Erkennung rhetorischer Sprachmittel. Aufgrund der zunehmenden Allgegenwärtigkeit von NLP-Systemen sind ethische Bedenken ebenfalls zu berücksichtigen.

Nach einer Einführung in NLP lag der zweite Beitrag dieser Ausarbeitung in der Betrachtung automatischer Textzusammenfassungssysteme und damit verbundener Herausforderungen. Textzusammenfassungssysteme wurden anhand folgender Kriterien kategorisiert:

- Eingabetyp – Zusammenfassung eines Textes oder mehrerer Texte
- Verwendungszweck – domänenspezifisch, generisch oder query-basiert
- Ausgabetyp – Sammlung von Sätzen des originalen Textes (*Extract*) oder Erstellung eines eigenen Textes (*Abstract*)

Zur *Extract Summarization* kommen überwiegend statistische Verfahren zum Einsatz. *Abstract Summarizations* sind hingegen aufgrund der Vielzahl an zu erledigenden Verarbeitungsaufgaben deutlich schwieriger zu realisieren. Hierfür haben sich neuronale Netze, insbesondere *Recurrent Neural Networks*, als nachhaltige, wirksame Lösung erwiesen. Anwendungsdomänen, in denen Zusammenfassungssysteme eingesetzt werden, sind zum Beispiel Webseiten, Nachrichtenartikel oder Emails. Festgestellt wurde zum Schluss, dass der Stand von Textzusammenfassungssystemen von Herausforderungen gekennzeichnet ist, welche auch sämtliche NLP-Systeme auch betreffen.

[9] https://www.agolo.com/ Abruf: 22.01.19

[10] https://deepai.org/machine-learning-model/summarization Abruf: 22.01.19

Darüberhinausgehende Herausforderungen wurden jedoch im Kontext von Textzusammenfassungssystemen identifiziert.

Insgesamt trägt diese Ausarbeitung dazu bei, einen Einblick in die Grundlagen, Anwendungen und Herausforderungen von *Natural Language Processing* zu erhalten. Derzeit geht die Tendenz in Richtung *Deep Learning,* um NLP-Systeme zu realisieren. Durch aktuelle Forschungsarbeiten und erhebliche Fortschritte im Bereich des Verstehens natürlicher Sprachen gewinnt NLP immer mehr an Bedeutung. Dies lässt sich unter anderem an der Popularität der diversen NLP-Systeme von *Google* oder *Amazon* ablesen.

LITERATUR

[1] Allahyari, M., et al. 2017. Text Summarization Techniques: A Brief Survey. *arXiv Preprint arXiv:1707.02268*.

[2] Chopra, S., Auli, M., and Rush, A.M. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2016), pages 93–98.

[3] Edmundson, H.P., 1968.New methods in automatic extraction. *J ACM, v.16 n.2, 264–285*. [doi>10.1145/321510.321519].

[4] Erkan, G. and Radev, D. R. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR).22*, 1 (2004), 457–479.

[5] González-Ibáñez R., Muresan S., and Wacholder N., 2011. Identifying sarcasm in Twitter: a closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, June 19-24, Portland, Oregon.

[6] Joanisse, M. F. and McClelland, J. L. 2015. Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6, 235–247. doi:10.1002/wcs.1340.

[7] Joshi, A.K. 1991.Natural language processing. *Science*, 253 (5025), pp. 1242-1249.

[8] Jurafsky, D. and Martin, H. J. 2018. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Pearson. Prentice Hall, Third Edition draf*t.

[9] Kumar, Y. J., et al. 2016. A review on automatic text summarization approaches. *Journal of Computer Science,* 12(4), 178–190. ISSN 1549-3636. doi: 10.3844/jcssp.2016.178.190.

[10] Leidner, L. J. and Plachouras, V. 2017. Ethical by Design: Ethics Best Practices for Natural Language Processing. *In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing,* pages 30–40, Valencia, Spain. Association for Computational Linguistics.

[11] Liddy, E. 2001. Natural language processing, *2nd edn. Encyclopedia of Library and Information Science*. Marcel Decker.

[12] Lloret, E. and Palomar, M. 2010. Challenging issues of automatic summarization: relevance detection and quality-based evaluation. *Informatica. Special Issue on Computational. Linguistics 34 (1):* 29–35.

[13] McCargar, V. 2005. Statistical Approaches to Automatic Text Summarization. *Bull Am Soc Inf Sci Technol 30(4):21-25*.

[14] Nadkarni P.M., Ohno-Machado L., and Chapman W.W. 2011. Natural language processing: an introduction. *J Am Med Inform Assoc.*, vol. 18 5(pg. 544-551).

[15] Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *In Proceedings of CoNLL*.

[16] Nenkova, A. and McKeown, K. 2012. A survey of text summarization techniques. *In Aggarwal, C. C. and Zhai, C., editors, Mining Text Data*.

[17] Raskin, V. 1985. Linguistics and Natural Language Processing. *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* Colgate University, Hamilton, NY, August, pp. 268--282. Auch, in: *S. Nirenburg (ed.), Machine Translation: Theoretical and Methodological Issues, ACL Series 'Studies in Natural Language Processing,* Cambridge University Press, pp. 42--58.

[18] Rush, A.M., Chopra, S., and Weston, J. 2015. A neural attention model for abstractive sentence summarization. *In EMNLP*.

[19] Saggion, H. and Poibeau T., 2013. Automatic Text Summarization: Past, Present and Future. *In: Multi-Source, Multilingual Information Extraction and Summarization, Poibeau, T., H. Saggion, J. Piskorski and R. Yangarber (Eds.), Springer Science and Business Media, Berlin*, ISBN-10: 3642285694, pp: 3-21.

[20] Tu, J.V.1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 49(11):1225–1231.

[21] Wermter, S., Riloff, E., and Scheler, G. 1996. Learning Approaches for Natural Language Processing. *In Connectionist Statistical and Symbolic Approach to Learning for Natural Language Processing*, Berlin: Springler.

[22] Westermann, G., Ruh, N., and Plunkett, K. 2009. Connectionist approaches to language learning. *Linguistics*, 47 (2). pp. 413–452. doi: 10.1515/LING.2009.015.