Grundlagen zu Natural Language Processing und Analyse von GATE als NLP-Tool

Aaron Schul, Felix Ritter

18. August 2018

Inhaltsverzeichnis

1 Einleitung

Mit voranschreitender Globalisierung und immer größeren Mengen an übertragenen Informationen steigt auch die Menge an zu verarbeitender natürlicher Sprache. Während dies bisher manuell durch den Menschen geschah, kommt allmählich aufgrund der schieren Menge von Daten vor allem im unternehmerischen Kontext und im Internet die Notwendigkeit auf, die natürliche Sprache automatisiert durch Computer verarbeiten zu lassen.

Dies bezeichnet man als Natural Language Processing (kurz NLP). Es hilft dabei, Texte beispielsweise nach Schlagworten zu durchsuchen, strukturell zu analysieren, Muster zu erkennen und teilweise sogar die Bedeutung von Geschriebenem zu verstehen und diese Semantiken darzustellen.

Mit dem Aufkommen von NLP steigt auch das Interesse an NLP-Tools, also an Programmen, welche in der Lage sind, die für NLP notwendigen Funktionalitäten übersichtlich und praktikabel bereitzustellen. Eines der bekanntesten NLP-Tools ist das von der University Of Stanford entwickelte GATE, welches eine Vielzahl von Anwendungsbereichen im NLP abdeckt.

Das Ziel dieser Arbeit ist es, eine Anforderungsanalyse für solche Tools durchzuführen und diese dann mit dem GATE-Tool abzugleichen. Somit bietet sich die Möglichkeit GATE zu analysieren und daraufhin kriterienbasiert zu evaluieren. Funktionale wie nicht-funktionale Anforderungen werden definiert.

Zu diesem Zweck werden zunächst ausführlich die Grundlagen zu NLP und dessen Tools erläutert. Die Teilschritte der Textverarbeitung werden theoretisch wie praktisch anhand von Implementierungsbeispielen erläutert. Daraufhin können mit Hilfe von Use Cases Anforderung an diese Tools erhoben werden, auf denen Analyse und Evaluation des GATE-Tools beruhen sollen. Abgeschlossen wird die Arbeit von einer kurzen Zusammenfassung der gewonnenen Erkenntnisse, gefolgt von einem Ausblick auf mögliche zukünftige Arbeiten und Entwicklungen für NLP-relevante Domänen.

2 Grundlagen von Natural Language Processing

2.1 Einführung

Der folgende Abschnitt befasst sich mit den Grundlagen zu Natural Language Processing (im folgenden kurz NLP). Dazu gehört neben dessen Herkunft bzw. Motivation eine inhaltliche Wissensgrundlage, welche für den weiteren Verlauf der Ausarbeitung relevant ist.

2.1.1 Motivation und Herkunft

Ein Großteil der menschlichen Kommunikation findet durch natürliche Sprache statt. Dies ist bereits seit Jahrtausenden so, jedoch hat sich die Übertragung dieser Sprache mit der Zeit verändert und weiterentwickelt.

Damals lediglich von Mund zu Mund übertragen war der erste große Schritt die Entwicklung der Schrift. Ob mit Hieroglyphen, Alphabeten oder anderen Zeichen, war man in nun in der Lage natürliche Sprache über lange Zeit und/oder weite Strecken zu vermitteln. Dies erwies sich als sehr Vorteilhaft und so setzte sich die Entwicklung fort, bis man über das Morsen und das Telefon schließlich die elektronische Nachricht erfand. Im Rahmen der voranschreitenden Digitalisierung und Globalsierung stieg die Menge an übertragener natürlicher Sprache weiter rasant an, sodass heutzutage ein Leben ohne beispielsweise die E-Mail unvorstellbar scheint. Obgleich die Übertragung durch kabelgebundene oder drahtlose Kommunikation weitgehend automatisiert ist, erfolgt die Auswertung des Inhalts weiterhin manuell durch menschliche Empfänger.

Die Daten selbst sind jedoch inzwischen kaum noch nur durch den Menschen effizient zu verarbeiten, sodass man nach einer neuen Möglichkeit sucht, dem Menschen diese Arbeit zu erleichtern, oder sogar abzunehmen. Mit diese Aufgabe beschäftigt sich NLP.

Bereits in den 1960er Jahren versuchte man durch Sprachcomputer bzw. Chatbots wie ELIZA (Weizenbaum 1966) die Mensch-Maschine-Kommunikation umzusetzen, indem die Textnachricht eines Benutzers automatisch durch einen Computer verarbeitet wurde und eine plausible Antwort gegeben wurde, damals jedoch noch ohne echte Wissensbasis. Heutige Entwicklungen im Bereich Sprachassistenzsysteme sind alltäglicher Begleiter jedes Smartphone-

Nutzers geworden; sie analysieren die Spracheingaben mithilfe von NLP-Techniken bezüglich ihrer Bedeutung in Echtzeit über Clouds.

2.1.2 NLP als Annäherung an natürlichsprachliche Probleme

Bereits bei der Mensch-Mensch-Kommunikation sind Verständigungsprobleme vorhandenen und Teil der Kommunikation selbst. Die Intention einer Aussage ist etwa nicht nur abhängig von dem, was tatsächlich gesagt wird, sondern auch etwa von Gestik, Mimik und der Situation, in der kommuniziert wird. Bei der Untersuchung von geschriebenem Text in natürlicher Sprache treten diese Betrachtungen jedoch in den Hintergrund, da sich auf den Inhalt konzentriert wird und mitunter die Situation des Autoren unklar oder irrelevant ist. Wissen über die Problemdomäne ist jedoch zwingend für das Verständnis spezifischer Fachbegriffe von Nöten, daher müssen auch automatische NLP-Tools diese einbeziehen können. Zudem gibt es oft Veränderungen der Bedeutung von Sprache die stark kontextabhängig sind, wie zum Beispiel Sarkasmus oder Ironie. Hier soll nicht weiter auf Aspekte der Kommunikationswissenschaften eingegangen werden, jedoch ist beispielsweise die Ambiguität einer Aussage ganz alltäglich und hat gleichsam verschiedene Auswirkungen. Solche Mehrdeutigkeiten werden durch syntaktische und semantische Fehlinterpretationen verursacht.

Wo sich Menschen im Zweifel auf Erfahrungen und spezifische Fachkenntnisse verlassen oder bei ihrem Kommunikationspartner nachfragen können, müssen sich Computer allein auf die vorliegenden Dokumente in Schriftform verlassen; sie wissen nichts über den Autoren oder dessen Blickwinkel. Wie später gezeigt wird, kann jedoch etwa Fachwissen durch domänenspezifische Wissensbasen simuliert werden.

Die Frage nach einer intelligenten, algorithmen- und / oder wissensbasierten Auswertung durch Computer stellt sich im Angesicht der aufgezeigten Besonderheiten natürlicher Sprache. Es gilt also für NLP nicht nur die Probleme zu lösen, die generell mit präziser Automatisierung verbunden sind, sondern zusätzlich so zuverlässig wie möglich die oben genannten und der Sprache inhärenten Komplikationen zu bewältigen. Bereits einfache Anfragen können computergestützte Frage-Antwort-Systeme wie SQL-Datenbanken an ihre Grenzen bringen, wie die verschiedenen Formen von Mehrdeutigkeiten zeigen können:

Schon das Wort *überblicken* kann das Übersehen von etwas, ebenso wie das im Blick haben (Gegenteil) bedeuten.

Der Satz *Die Betrachtung des Studenten* kann etwa als Student der (etwas) betrachtet, oder als ein Student, der (von jemand anderem) betrachtet wird, verstanden werden. Sätze mit solch einem Satzbau (bestehend aus nominalisiertem Verb, Bezugswort und Substantiv) verursachen unwillkürlich zwei mögliche Deutungen aufgrund der unklaren Syntax.

Die Frage *Verkaufen Sie Handys und Computer von Samsung?* kann als Frage nach allgemeinen Handys und nur Computern, spezifisch von Samsung oder nach Handys und Computern gleichermaßen nur von Samsung verstanden werden. Auch hier entsteht Unklarheit durch zwei mögliche Bezüge des einschränkenden Relativsatzes. Ob eine Aufzählung oder ein Unterschied (zwischen *von Samsung* und *nicht von Samsung* gemacht wird, ist syntaktisch unklar.)

Wie schnell ist der Bus? und *Wie schnell ist der Bus da?* mögen zunächst ähnlich aussehen, fragen jedoch inhaltlich anders und haben nichts miteinander zu tun. Außerdem könnte der zweite Satz kontextabhängig etwa nach der Fahrzeit bis zum Ziel oder nach der Ankunftszeit, bis an der man einsteigen kann, fragen. Die vage Formulierung von Sätzen stellt bei der Frage nach deren Bedeutung eine Herausforderung dar, durch Untersuchung des Kontextes der Frage kann hier jedoch meist recht treffsicher geantwortet werden.

Abgesehen von diesen syntaktischen Uneindeutigkeiten muss NLP auch Hürden der inhaltlichen Mehrdeutigkeit bzw. Semantik bewältigen. Aktuell beschäftigt man sich zum Beispiel mit der Aufgabe zu erkennen, ob ein Kommentar zu einem Video der Webseite Youtube eher positiv oder eher negativ gesinnt ist. Was hier für den Menschen schon beim ersten Lesen sofort erkennbar wird, stellt für den Computer ein ernstzunehmendes Problem dar. Tonalität und Stimmung des Kommentars, der unter einem Video als Reaktion entstehet, ist mitunter nicht direkt aus dem audiovisuellen Eindruck aus dem Video ersichtlich. Solche scheinabr simplen Probleme sind für den Menschen scheinbar simpel zu lösen. Wie oben erwähnt existiert NLP bereits seit ca. Mitte des 20. Jahrhunderts, jedoch stellt die Behandlung natürlichsprachlicher Probleme auch heute noch eine Herausforderung bei der Implementierung dar. Wenn Systeme die oben genannte Beispiele nicht explizit ausschließen, muss die Verarbeitungslogik besondere Rücksicht auf Uneindeutigkeiten nehmen.

2.1.3 Definition

NLP ist grob definiert als die automatische oder halb-automatische Verarbeitung von natürlicher Sprache. Manche schließen aus dieser Definition die Erfassung der Sprache aus, da diese nicht Teil der eigentlichen Verarbeitung ist. (Der Einfachheit halber wird im weiteren Verlauf davon ausgegangen, dass die Sprache als digitale Textdatei vorliegt, wenn nicht explizit anders angegeben. Es kann sich jedoch auch um gesprochenes Wort oder Gesten handeln)

NLP überschneitet sich mit vielen wichtigen Bereichen der Wissenschaft. Hauptsächlich sind dies Linguistik und Informatik, allerdings fließen auch Psychologie, Philosopie und Mathematik bzw. Logik stark mit ein.

2.1.4 Ziel und funktionale Einordnung

Das Ziel von NLP ist die Extraktion von Informationen aus gegebenen Texten, also aus einem natürlichsprachigen Input einen Wissensoutput über den Inhalt des Dokumentes zu generieren. Dies geschieht praktisch mithilfe von NLP-Tools wie GATE.

Dazu ist, wie im nächsten Abschnitt beschrieben, die linguistische Analyse der Dokumente der Sprachstruktur in Teilschritten erforderlich. Auf Basis der syntaktischen Analyse kann dann eine semantische Analyse durch Einsatz sogenannter Ontologien (Sammlung von Wissen aus einer Domäne) erfolgen. Letztendlich kann somit tatsächlich Wissen über die Bedeutung des Dokumenteninhalts gewonnen werden.

2.2 Linguistische Analyse

Der folgende Abschnitt beschäftigt sich mit der Analyse der sprachlichen Struktur eines Dokumentes. Am Anfang von NLP steht die linguistische Analyse des Textes etwa bezüglich Satzstruktur und Differenzierung von Wörtern aus der Sprache und Eigennamen. Das Ziel ist die sogenannte Annotation der Textbausteine bezüglich ihrer Bedeutung. Sachzusammenhänge und Relationen von Wörtern sollen dargestellt werden.

NLP kann, wie später in dieser Arbeit beschrieben, als Zusammenfassung von 6 wesentlichen Teilbereichen verstanden werden. Begonnen bei der einfachen Wortanalyse, werden später Bedeutungen und Verknüpfungen zur Anwendbarkeit der Aussage eines Satzes oder Wortes in der realen Welt

analysiert. Die Linguistik befasst sich seit jeher mit der Untersuchung von natürlicher Sprache im Hinblick auf die dahinterliegenden Konstrukte. NLP spaltet bei der Analyse der Sprachbestandteile weiter auf, etwa müssen Die folgenden Teilbereiche, die daraus bekannt sind, sind bei der natürlichsprachigen Analyse identifiziert worden:

1. Morphologische Analyse - Die Zerlegung von Wörtern bezüglich ihrer Struktur. Die Komposition aus Präfix, Wortstamm und Suffix von Wörtern wird erkannt, um Fall, Tempus, Numerus etc. des Wortes zu bestimmen. Im Englischen zeigt so die Endung -ed bei Verben die Vergangenheitsform an. Dabei ist zu beachten, dass Mehrdeutigkeiten entstehen können.

2. Syntax - Zuordnung von Bedeutungen anhand sprachlicher Phrasen. Jede Sprache besitzt syntaktische Regelungen bezüglich der auftauchenden Wörter; im Deutschen folgt etwa auf einen Artikel stets ein Nomen oder bestimmte Signalwörter geben Satztyp und Tempus an. Auf Basis dieses Wissens können formale und strukturelle Analysen der Textbestandteile erfolgen.

3. Semantiken - Die Identifikation einer Bedeutung von Sätzen und Begriffen. Die Semantik eines Textabschnittes wird häufig als "Logik"bezeichnet und fragt nach dessen Bedeutung bzw. Thema. Bedeutungen können anhand von Kompositionen einzelner Wörter und Sätze auf Basis der semantischen Analyse erkannt werden.

4. Kontext - Darstellung von Sachzusammenhängen aus der Vereinigung von ähnlichen Bedeutungen. Textbausteine, die sich mit dem gleichen Thema beschäftigen, werden dem gleichen Sachzusammenhang zugeordnet und als Teil dessen annotiert.

2.3 Allgemeine NLP Architektur

Die Verarbeitung eines Dokumentes mittels NLP erfolgt nacheinander in einzelnen Schritten, die jeweils einen Bereich oder Teilbereich der Linguistik des Textes abbilden. Die Analyse erfolgt dabei nacheinander und wird immer feingranularer, begonnen bei einfachen grammatikalischen Analysen bis hin zu komplexen Wissensbasierten verfahren. Die folgenden Schritte stellen laut EINIGEN QUELLEN eine typische komponentenweise Architektur dar.

Es ist anzumerken, dass auch mehrere Schritte gemeinsam in einem Schritt durchgeführt oder gar ganz entfallen können, abhängig von der Umsetzung anderer Stufen:

1. Input Processing - Erkennung der Dokumentensprache und Normalisierung des Textes. Im ersten Schritt geht es um das korrekte Format des Eingabedokumentes für die Verarbeitung. Dieser Vorgang stellt jedoch noch keine Analyse dar, sondern dient lediglich der Vorbereitung.

2. Morphologische Analyse - Die meisten Sprachen sind im Bezug auf ihre Grammatik und syntaktische Struktur der Wörter in Systemen abbildbar, etwa durch Modellierung mittels Automaten zur Erzeugung der Worte. Auch können Vokabeln in Wörterbüchern/Lexika gesammelt und Regeln auf ihnen formuliert und gespeichert werden.

3. Part-of-Speech-Tagging - Die einzelnen Wörter eines Satzes werden im Hinblick auf den Sach- oder Satzzusammenhang, wie auch auf die Stellung der Stellung im Satz hin analysiert. Basierend auf Kontext und/oder Erfahrungswerten bzw. Heuristiken können die Wörter korrekt erfasst und deren Fall getaggt werden. Subjekte, Prädikate usw. werden identifiziert. Dazu ist eine Wissensbasis mit Trainingsdaten und die Einordnung des Kontextes darin von Nöten.

4. Parsing - Die Ergebnisse der der vorherigen Schritte werden weiter verarbeitet und in ein standardisiertes Format gebracht. Syntaktische Zusammenhänge, wie etwa das Zusammenfassen von Wörtern zu einer gemeinsamen Bedeutungsphrase und das Zuordnen von Verben zu einem Nomen können nach dem Parsing dargestellt werden.

5. Disambiguation - Die Entfernung von Mehrdeutigkeiten auf Basis der Ergebnisse des Parsings stellt einen entscheidenden Schritt für die semantischen Analysen dar. Nur wenn die Aussage und der Kontext eines Satzes klar erkennbar sind, kann dessen Semantik gezielt abgeleitet werden.

6. Context Module - Textbausteine, deren Interpretation semantisch auf dem Kontext anderer Sätze oder Wörtern beruhen, können erfasst werden. Bei Anaphern ist etwa das Verständnis des aktuellen Kontext abhängig von einer vorherigen Deutung.

7. Text Planning - Die Sachzusammenhänge und Semantiken, die aus dem zugrundeliegenden Text extrahiert wurden, werden für die Darstellung im fertigen Text definiert. Es wird festgelegt, welche der Bedeutungen übertragen und vermittelt werden sollen.

8. Tactical Generation - Die Bedeutungen werden in Form konkreter Zeichenketten generiert, die die gewünschte Bedeutung enthalten. Diese Textbausteine basieren häufig direkt auf den Ergebnissen des vorherigen Parsings, da dort schon Bedeutungen zusammengefasst werden können.

9. Morpholigical Generation - Die erzeugten Sätze werden morphologisch an die Rahmenbedingungen der verwendeten Sprache angepasst. Grammatiken und Regeln der Satzkonstruktion werden auf die erzeugten Wörter angewendet, sodass ein lesbarer und nachvollziehbarer Text entsteht.

2.4 Morpholoie in NLP

Der folgende Abschnitt befasst sich genauer mit der morphologischen Analyse von Wörtern und den dazu verwendeten Automaten. Da zu der englischen Sprache bereits die meisten Fortschritte erziehlt wurden und sie verhälnismäßig unkomplex im Vergleich zu anderen Sprachen ist, werden im Folgenden viele der Beispiele auf Englisch behandelt.

Wie bereits erwähnt befasst sich die Morphologie mit der Struktur, Zusammensetzung und Korrektheit einzelner Wörter. Jedes Wort der Englischen sprache besteht aus mindestens einem Wortstamm und beliebig vielen Affixen. Affixe sind unterteilt in Präfixe und Suffixe, abhängig davon, ob sie vor oder hinter dem Wortstamm auftauchen. Es gibt Affixe, welche auch in der Mitte des Wortes eingeordnet werden. Diese existieren jedoch nicht in der englischen Sprache und sind daher künftig zu vernachlässigen.

Man betrachte beispielsweise das Wort "believe"von dem englischen Verb "to believe"(zu Deutsch: "glauben"). Hängt man das Suffix ableän so erhält man das Adjektiv "believable"("glaubhaft"), fügt man nun das Präfix ün-"hinzu erhält man die Negation ünbelievable"(ünglaubhaft"). Die morphologische Analyse erkennt dies nicht nur als ein Wort, sondern als Präfix-Stamm-Suffix-Konstrukt.

Es wird unterschieden zwischen ableitender und flexionaler Morphologie, wobei die Grenze jedoch nicht ganz einfach zu setzen ist. Affixe wie änti-",

"re-ünd das oben verwendete ün-"gelten als ableitend und können mehrfach und sogar rekursiv zu Stämmen auftauchen. Sie fügen in der Regel eine Zusatzinformation an den Stamm an, wie z.B. Negation. Es kann auch, wie in dem Beispiel oben demonstriert, zur Änderung der Wortart kommen. Flexionale Affixe werden normalerweise wegen grammatikalischer Regeln wie Pluralisierung hinzugefügt. Hierzu gehören das Plural-sünd die Endung edïn der Vergangenheitsform.

2.4.1 Buchstabierregeln mit endlichen Automaten

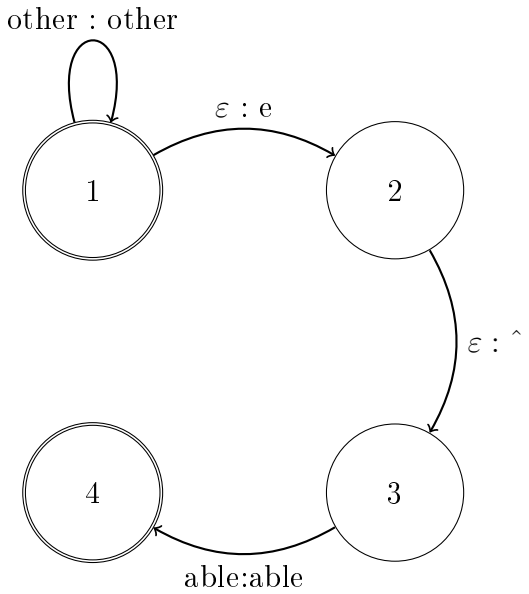
In der Morphologie gibt es eine Vielzahl von Regeln die bei dem Unterteilen, Interpretieren und Zusammensetzen von Wörtern zu beachten sind. Wenn man das "believeBeispiel aus dem obigen Abschnitt erneut betrachtet, fällt auf, dass bei dem hinzufügen des Suffixes able"der letzte Buchstabe des Wortstammes entfällt. Dies ist natürlich kein Tippfehler, sondern beruht auf einer der besagten Regeln. Die oben verwendete Regel lässt sich wie folgt darstellen:

$$\mathrm{e} \rightarrow \varepsilon \hat{} \_\mathrm{able}$$

Das e ist in dem Falle der in Frage stehende Buchstabe, welcher, symbolisiert durch den Pfeil, in das ε (das leere Wort) gewandelt wird, falls ein Affix angehängt wird (ˆ) und es an dieser Stelle (_) steht, also vor dem äble".

Solche Regeln lassen sich zu morphologischen Zwecken gut mit endlichen Zustandswandlern (finite state transducers) realisieren, was im Folgenden demonstriert werden soll.

Ein endlicher Zustandswandler ist ein Graph aus einer endlichen Anzahl von Knoten (Zuständen), welche durch Kanten verbunden sind. Knoten sind unterteilt in einen Start-Knoten, eine beliebige Anzahl an normalen Knoten und mindestens einen Ziel-Knoten. Der Start-Knoten als Input das zu verarbeitende Wort, welches nun in eine Abbildung der Aufbaustruktur des Wortes übertragen werden soll. Das Erreichen eines Ziel-Knotens heißt, dass das Wort korrekt ist und akzeptiert werden kann. Die Kanten repräsentieren die Abbildung einer Zeichenkette auf eine andere. Dies funktioniert bidirektional, wodurch sie sowohl zum Erstellen, als auch zum Auflösen von Wörtern verwendetet werden können. Die folgende Abbildung zeigt den eindlichen Zustandswandler, der die Umsetzung der besagten äbleAnhangs-Regel ermöglicht:



Der oben abgebildete Zustandswandler ist in der Lage die beschriebene Regel umzusetzen wie folgt: In Zustand 1 erfolgt der Input des eingelesenen Wortes, also in diesem Falle "believable". Über die Schleife mit öther:otherïst die Abbildung aller nicht speziell angegebenen Zeichenfolgen auf sich selbst beschrieben. Es kann jederzeit das leere Wort auf ein ë"Übertragen werden, dies geschieht jedoch nur, falls auch ein äble"folgt. In dem Fall wird zuvor noch aus dem leeren Wort ein Affix-Zeichen (^) generiert (Es ist anzumerken, dass der Zustandswandler nicht deterministisch ist, sich jedoch über die Potenzmengenkonstruktion immer ein äquivalenter deterministischer Zustandswandler erstellen ließe). So wird also das Wort "believable"durch diesen Automaten als "believe^ableÏnterpretiert, und somit das äbleAffix erkannt. Andersherum ließe sich aus dem Wortaufbau "believe^able"das korrekte Wort "believable"generieren.

2.5 Part of speech tagging und Wortvorhersage

Dieses Kapitel befasst sich mit der Transformation der Ergebnisse der morphologischen Analyse einzelner Wörter in eine sinnvolle Kategorisierung der

Wörter im Sachzusammenhang aus mehreren Wörtern. Part-of-speech tagging klassifiziert einzelne Worte eines Textes so, dass erfasst wird, welchen Teil der Sprache sie ausmachen. Die einzelnen Wörter des Textes werden bezüglich ihrer grammatischen Funktion und Anzahl markiert (*annotiert*). Im Deutschen und Englischen sind die grammatikalischen Regeln im Verhältnis zu etwa chinesisch oder türkisch recht einfach erfassbar, da Satzbau und Wortbildung die Wortfunktion definieren. Noam Chomsky befasste sich sogar mit der Frage, wie sich das Englische formalisieren, das heißt etwa mit Hilfe von Automaten modellieren, lässt. Die Spracherfassung natürlicher Sprache durch Computer wurde somit ermöglicht. Die Identifikation einzelner Bausteine muss derart erfolgen, dass sie den Kontext des Satzes erfasst, also Sinn ergibt. Diese Kennzeichnung geht über die Analyse einzelner Wörter hinaus, sie erfasst auch die Syntax umgebender Wörter eines Satzes.

Im Deutschen ist es etwa recht wahrscheinlich, dass auf ein Nomen am Satzanfang als nächstes Wort ein Verb folgt, da dies einen typischen Satzbau darstellt. Ein part-of-speech tagger sollte daher beispielsweise im Satz *Ïch fahre Fahrrad."* korrekt das Ïchäls Nomen, das "fahreäls Verb und das "Fahrradäls Akkusativobjekt darstellen; Die isolierte "Betrachtung von "Fahrrad"ließe aber auch die Kategorisierung als Nomen zu; diese Mehrdeutigkeit einzelner Worte im Satz wird durch Analyse des syntaktischen Kontextes eliminiert.

Solche syntaktischen Regeln aus einer Sprache werden in Form von sogenannten *corpora* geliefert; sie stellen Trainingsdaten in Form von tatsächlicher Sprache bzw. Text dar. Im Englischen wird etwa häufig die Zeitung The Wall Street Journal genutzt, um die Repräsentation bestimmter Satzstrukturen und Wortfolgen statistisch zu erfassen. Alternativ können auch standardisierte Corpora wie der Lancaster-Oslo-Bergen Corpus verwendet werden, der etwa eine Million Wörter enthält. Corpora sind etwa für NLP auf Basis von Machine-Learning-Algorithmen erforderlich, können darüber hinaus aber auch etwa zur Rechtschreibprüfung oder Textunterteilung benutzt werden. Die meisten Textverarbeitungsprogramme können heutzutage etwa erkennen, wenn sich syntaktische Fehler in einem Satz ereignen (Kommasetzung, falscher Kasus, etc.). Es ist anzumerken, dass part-of-speech tagging nicht unbedingt vorher annotierten Texten bedarf, aber ohne diese fällt eine syntaktische Einordnung ungleich schwerer und die Fehlerquote steigt massiv. Mehrdeutigkeiten aus der Syntax heraus können somit nicht aufgeschlüsselt werden.

Die Anwendung eines Corpus auf ein natürlichsprachliches Dokument er-

möglicht sogleich auch eine Vorhersage der nächsten Wörter bzw. Sprachbausteine, nachdem schon ein Teilsatz gelesen wurde. Diese *prediction* kann auf Basis bekannter Regelmäßigkeiten also schon vor der Wortanalyse recht genau die nächste grammatikalische Komponente bzw. den Worttyp vorhersagen.

2.5.1 N-gram Modelle zur Sprachvorhersage und -modellierung

Auf Basis der vorliegenden Corpora existieren eine Vielzahl an Regeln bezüglich des Satzbaus und bekannnter Folgen von Worttypen, wenn ein neues Dokument annotiert werden soll. Zur Veranschaulichung einer möglichen Annotation soll in diesem Abschnitt die Notation von CLAWS 5 (QUELLE) für Annotationen genutzt werden.

Bei erneuter Betrachtung des obigen Beispiels *Ïch fahre Fahrrad."* könnte, wenn dieser Satz als minimaler Corpus vorliegt, neben der Regel Subjekt-Prädikat-Objekt eines Satzbaus auch beinhalten, dass der Punkt ein Satzende markiert und das Substantive und Personalpronomen mit einem Großbuchstaben beginnen (Ïchünd "Fahrrad"können Nomen sein und sind groß geschrieben). Sprachvorhersage nutzt diese bekannten Regeln und würde etwa aus der obigen Regel herleiten, das, wenn ein Verb gelesen wurde, darauf immer ein Substantiv bzw. das Satzobjekt folgt. Kontextfreie Bertrachtung einzelner Wörter, also ohne andere Taggings mit einzubeziehen, wird als Unigram bezeichnet. Dabei werden jedoch Regeln über den Zusammenhang von Wortketten vollständig ignoriert und das Ergebnis ist ähnlich dem der morphologischen Analyse. Vorhersage-Systeme, die sich der Annotierung des vorigen Wortes bedienen, werden als Bigramme (englisch *bigrams*) bezeichnet, Trigrams dementsprechend bei Berücksichtigung der zwei vorigen Wörter und N-gramme, wenn alle n-1 zuvor gelesenen Wörter, also die gesamte Terminologie des Textes bis zum n-ten Wort, berücksichtigt werden. Die Wahrscheinlichkeit des Typen des Folgewortes auf Basis eines Bigrams lässt sich berechnen, indem aus einem Corpus die Worttypen an der jeweiligen Satzstelle gezählt und statistisch erfasst werden. Aus folgendem Corpus vierer Aussagen lässt sich dies wie folgt ableiten (zur Vereinfachung werden hier tatsächliche Wörter und nicht deren Annotation getaggt):

(i)ich mag Züge (ii)ich mag guten Tee (iii)guten Morgen (iv)guten Abend

Zwischen einzelnen Äüßerungen wird der Platzhalter <s> als Indikator ei-

ner neuen Aussage verwendet, sodass dem Bigram nun eine Zeichenkette mit den obigen Aussagen, getrennt durch <s> zur Verfügung steht. Berechnet werden nun die Wahrscheinlichkeiten des Folgewortes auf Basis der vorherigen, formal ausgedrückt: $\frac{C(w_{n-1}w_n)}{\sum wC(w_{n-1}w}$

Sequenz	Anzahl	Wahrscheinlichkeit

Anhand des folgenden Corpus ergibt sich etwa:

Literatur