Universität Hamburg
Department Informatik
Knowledge Technology, WTM

Erfassen natürlicher Sprache in Textform am Beispiel von Python und C++

Proseminararbeit

Proseminar: Künstliche Intelligenz

Louis Kobras
Matr.Nr. 6658699
4kobras@informatik.uni-hamburg.de

30. Juni 2015

Abstract

Seit Jahrzehnten schon beschäftigen sich Forscher damit, den Computer dem Menschen näher zu bringen. Ein wichtiger Bestandteil dieses Prozesses ist dabei die Sprachverarbeitung.
Ziel dieser Arbeit ist, Licht auf den Prozess des Verständnisses eines Computers von natürlicher Sprache zu werfen.
Zu diesem Zweck wird der Begriff der natürlichen Sprache in seine Komponenten zerlegt, die einzelnd analysiert und in dieser Form von nichtkomplexen Automaten verarbeitet werden können.
Einbindung der Ergebnisse.

Inhaltsverzeichnis

1 Einführung **2**
- 1.1 Was bisher geschah: Sprachverarbeitung im Kontext der historischen Artificial Intelligence . . . 2
- 1.2 ELIZA als Beispiel für frühe Sprachverarbeitung . . . 3

2 Was ist natürliche Sprache? **3**

3 Schrittweise Verarbeitung natürlicher Sprache **4**
- 3.1 Parsing der Morphologie . . . 5
- 3.2 Analyse der Syntax . . . 5
- 3.3 Interpretation der Semantik . . . 5
- 3.4 Erfassung der Pragmatik . . . 5
- 3.5 Inbetrachtziehen des Diskurs . . . 5
- 3.6 Auflösung von Mehrdeutigkeit . . . 5

4 Simon, Sirius, Jasper **5**
- 4.1 Simon listens . . . 5
- 4.2 Sirius . . . 5
- 4.3 The Jasper Project . . . 5

5 Über Haushaltsroboter bis C3PO **5**
- 5.1 Das intelligente Haus . . . 5
- 5.2 Rechercheunterstützende Computer . . . 6
- 5.3 R2D2 und C3PO . . . 6

6 Schlussfolgerung **7**

Quellenverzeichnis **8**

Was gestrichen wird, sollte der Text zu lang werden (in dieser Reihenfolge, ggf. noch mehr):

- 1.1 und 1.3
- 5.1
- 4.1

1 Einführung

„Ich wünschte, du würdest das nicht tun. Was ist, wenn ein integrierter Schaltkreis versagt und ich nicht rechtzeitig eingreifen kann?“
„Ach, Andromeda, das würde dir ja nie passieren, denn dann brauchst du ja einen neuen Captain.“

*- Andromeda und Dylan Hunt, [**?**]*

Was wir hier vor Augen haben ist der erste Dialog zwischen der KI des Raumschiffes *Andromeda Ascendant* und ihrem Captain, den die Zuschauer der Serie von Gene Roddenberry zu hören bekommen. Der Captain springt einen tiefen Schacht herunter, um kurz vor Aufschlag mit einem Raketengürtel seinen Fall abzubremsen.
Ein physikalisch interessantes Szenario, doch warum wird es hier eingebracht?
Die KI dieses Raumschiffes, die im Folgenden nach ihrem androiden Avatar *Rommie* bezeichnet wird, ist in der Lage, mit den Besatzungsmitgliedern des Schiffes durch direkte, natürliche Sprache und das Ausdrücken von Emotionen zu kommunizieren. Dies ist zweifelsohne ein erstrebenswertes Ziel der Forschung an der Künstlichen Intelligenz. Doch wo es Ziele gibt, gibt es auch Wege zum Ziel. Diese Arbeit wird sich mit dem Verarbeiten natürlicher Sprache in Textform beschäftigen. Zwar ist Kommunikation durch Texteingabe nicht so schnell und dynamisch anwendbar wie gesprochene Anweisungen, dennoch aber ist sie ein wichtiger Schritt dorthin. Wir werden sehen, wie natürliche Sprache aufgebaut ist, wie sie durch Verwendung von Deterministischen Automaten und Markov-Ketten analysiert werden kann, und wie mit den erhaltenen Informationen verfahren werden kann.

1.1 Was bisher geschah: Sprachverarbeitung im Kontext der historischen Artificial Intelligence

Die Grundsteine für Künstliche Intelligenz wurden 1936 von Turing und darauf aufbauend in den 40ern und 50ern gelegt (McCulloch und Pitts (1943), Chomsky (1956) et. al.).
Ende der 50er Jahre hatten sich daraus zwei parallele Zweige für die Sprachverarbeitung gebildet: Das formale Paradigma und der stochastische Ansatz ([2]).
Während das formale Paradigma seinen Schwerpunkt auf formale Sprachen und das Parsen von Schlüsselwörtern gelegt hatte, lag der Schwerpunkt des anderen Ansatzes auf der Berechnung der Wahrscheinlichkeit des Auftretens bestimmter Wörter anhand

eines umfassenden Lexikons und der Anwendung Bayesscher Methoden auf die einzelnen Zeichen.
Während der stochastische Zweig weitestgehend erhalten blieb, teilte sich der formale Zweig zu Beginn der 70er Jahre in drei Pfade auf. Der stochastische Zweig begann, versteckte Markov-Modelle (HMM) bei seiner Analyse zu nutzen. Ein Logik-basierender Zweig befasste sich mit der Vereinheitlichung der Struktur von Sprache und nutzte dabei eine sog. *Definite Clause Grammar*, woraus sich später Prolog entwickeln würde. Der Zweig, der sich mit natürlicher Sprache befasste, versuchte, Text-Kommandos und Konzepte so umzusetzen, dass Maschinen eine semantische Grundlage mitgegeben werden konnte, auf der die Interpretation aufbauen würde. Der Diskurs-Zweig verwendete die Betrachtung von Substrukturen im Diskurs von Sprachen und der Automatisierung von Referenzen.
Die 80er und 90er Jahre beschäftigten sich Whahrscheinlichkeitsmodellen und Modellen, die sich deterministische, endliche Automaten zur Analyse zu Nutzen machten.
Ende des 20. Jahrhunderts hatte sich das Wahrscheinlichkeitsmodell als Standard durchgesetzt. Die Forschung in diesem Feld wurde durch das Aufkommen des Internet und auf kommerzieller Basis durch Entwicklung der Technologie weiter vorangetrieben.

1.2 ELIZA als Beispiel für frühe Sprachverarbeitung

ELIZA ist ein frühes Programm zur Sprachverarbeitung (Weizenbaum, 1966). Das Programm wendet simples Parsing und Mustererkennung an, um Schlüsselwörter aus einem gegebenen Satz zu extrahieren und darauf aufbauend eine Antwort zu generieren. ELIZA arbeitet als Rogativer Psychologe, d.h. man erzählt ihr etwas und sie stellt eine darauf basierende Frage.
Das Programm ist sehr simpel, und ELIZA benötigt fast keine Informationen, um zu funktionieren. Durch ihre Arbeitsweise kann sie ein natürliches Gespräch mit einem Psychologen nur durch das Wiederholen von Sätzen in Frageform erfolgreich simulieren.

2 Was ist natürliche Sprache?

„By 'natural language' we mean a language that is used for everyday communication by humans; languages such as English, Hindi, or Portuguese. In contrast to artificial languages such as programming languages and mathematical notations, natural languages have evolved as they pass from generation to generation. and are hard to pin down with explicit rules.“ [1]
So definieren Steven Bird et. al. den Begriff *natürliche Sprache*. Eine Sprache ist eine Konvention zur Kommunikation bestehend aus einer Grammatik, einem Vokabular und einem Verwendungskontext.
Eine natürliche Sprache ist eine Sprache, die sich im Verlauf der Zeit aus der Kommunikation zwischen Menschen entwickelt hat. Durch die permanente Entwicklung einer Sprache verändert sich permanent die Bedeutung und Verwendung von Wörtern, andere Wörter werden neu hinzugefügt, wieder andere fallen heraus. So wurde zum Beispiel

das Wort `Gebäudereinigungsfachkraft` erst vor relativ kurzer Zeit in den aktiven Wortschatz seiner Sprache aufgenommen, während das wort `archaistich` für 'veraltet' oder 'altertümlich' kaum noch Verwendung findet. Das erste Beispiel hat aus Gründen der politischen Korrektheit nicht nur Einzug in die deutsche Sprache als neue Bezeichnung gefunden, sondern auch seinen Vorgänger ersetzt.
Bird et. al. nahmen ebenfalls Bezug auf künstiche Sprachen. Programmiersprachen und mathematische Notationen sind, ebenso wie natürliche Sprachen, als Verbindung aus Grammatik, Wortschatz und Verwendung darstellbar. Der Unterschied zu den natürlichen Sprachen ist jedoch, dass diese Sprachen feststehen. In der Sprache der Mathematik findet außer in Fachkreisen kaum noch eine Veränderung statt. Ebenso verhält es sich mit dem täglich verwendeten Java. Mit jedem Update gibt es eine handvoll Änderungen, jedoch nimmt kaum eine dieser Änderungen Einfluss auf den alltäglichen Gebrauch der Sprache.
Die Eigenschaft von natürlichen Sprachen, dynamisch zu sein, macht es schwer, sie in kompakten Regeln klar zu formulieren. Um es dennoch zu können, wird eine Sprache in folgende sechs Teilgebiete aufgeteilt, die jedes für sich erfasst und bearbeitet werden kann: Morphologie, Syntax, Semantik, Pragmatik, Diskurs und Mehrdeutigkeit.
Für die gesprochene Sprache kommt der Punkt der Phonologie hinzu.
Unter **Morphologie** verstehen wir die Veränderung des Wortes nach Anwendung der Grammatik.
Unter **Syntax** verstehen wir [...]
Unter **Semantik** verstehen wir [...]
Unter **Pragmatik** verstehen wir [...]
Unter **Diskurs** verstehen wir [...]
Unter **Mehrdeutigkeit** verstehen wir, dass ein Wort verschiedene Bedeutungen abhängig vom Kontext haben kann. Unter **Phonologie** verstehen wir, dass sich je nach Aussprache die Bedeutung eines Wortes ändern kann.

3 Schrittweise Verarbeitung natürlicher Sprache

TODO:
tiefergehende Erläuterung der Begriffe und Begründung der gewählten Reihenfolge
hauptsächlich Eingehen auf [2]
wird Code-Beispiel und/oder Automaten enthalten

3.1 Parsing der Morphologie

3.2 Analyse der Syntax

3.3 Interpretation der Semantik

3.4 Erfassung der Pragmatik

3.5 Inbetrachtziehen des Diskurs

3.6 Auflösung von Mehrdeutigkeit

4 Simon, Sirius, Jasper

Diese drei Namen sind Namen von Open Source Sprachverarbeitungssoftware. Wir werden diese drei Pakete untersuchen und ihre Gemeinsamkeiten bezüglich der Sprachanalyse ermitteln.
- wird Code-Beispiele enthalten

4.1 Simon listens

KDE-basierendes Programm
Sprache: C++

4.2 Sirius

Für die meisten Linux-Distributionen verwendbar
Sprache: Python, C++, Shell

4.3 The Jasper Project

Software für Raspberry Pi
Sprache: Python

5 Über Haushaltsroboter bis C3PO

Es wurde gezeigt, wie Sprachverarbeitung ausgesehen hat (1) und wie das Verständnis natürlicher Sprache funktioniert (3). Doch was kommt nun? Was hat man davon, dass Computer Menschen verstehen können? An drei Beispielen aus der Fiktion werden nun mögliche Anwendungsgebiete gezeigt.

5.1 Das intelligente Haus

„Herzlich Willkommen.“ - „Was ist das?“ - „DAS war Sarah.“
„Selbstständig arbeitendes, rundum automatisiertes Haus. Kurz: S.A.R.A.H.“

*- Sarah, Jack Carter und Douglas Fargo, [**?**]*

Sarah ist ein Haus aus der Science Fiction-Serie EUREKA. Tatsächlich ist *Sarah* weniger Fiction als Science. Es, oder Sie, ist tatsächlich eine künstliche Intelligenz, die in ein Haus eingebaut wurde. *Sarah* verwaltet dabei sämtliche Funktionen des Hauses: Lüftung, Raumtemperatur, Küche, Multimedia, Türen. Sie kommuniziert mit den Bewohnern des Hauses durch gesprochene Sprache. So bittet der Hausherr sie zum Beispiel mit dem Aufruf *„Sarah, Tür!“* darum, die Tür zu öffnen oder zu schließen. Doch über rudimentäre Befehle hinaus können die Bewohner auch komplexe Unterhaltungen über den Beruf, die Schule oder die Gefühle führen. Nun wird viel nötig sein, um Empathie oder Emotionen simulieren zu können. Jedoch ist es nicht unrealistisch, davon auszugehen, dass in nicht allzu ferner Zukunft solche Gespräche zwischen Mensch und Maschine möglich sind.

5.2 Rechercheunterstützende Computer

„Computer, erbitte alle verfügbaren Informationen über die fiktive Figur 'Dixon Hill'.“ „Bitte warten. Die Figur tauchte zum ersten Mal in dem Magazin 'Erstaunliche Detektivgeschichten' auf, [...]“

*- Commander Data und der Schiffscomputer; [**?**]*

Recht früh in der Serie *Star Trek: The Next Generation* gibt es eine Folge, in der der Android Data dem Computer des Raumschiffes *Enterprise* eine Rechercheanweisung gibt. Der Computer nimmt einen Suchbegriff entgegen und scannt die komplette Datenbank der Föderation nach allen möglichen Treffern. Hierbei muss er differenzieren, was als möglicher Treffer und was als tatsächlicher Treffer gilt. Tatsächlich kann der Computer nur anhand dieser Anweisung nicht entscheiden, welcher *Dixon Hill* gemeint ist, sollten mehrere existieren. Hierbei müssen Wahrscheinlichkeiten verglichen und der Kontext analysiert werden. Dies in diesem Ausmaße in absehbarer Zeit serientauglich zu machen, halte ich für unwahrscheinlich. Dennoch, möglich wäre es.
(Der Computer hier hatte ein wenig Hilfe - Der Captain des Raumschiffes hatte kurz zuvor ein Unterhaltungsprogramm mit Fokus auf den Privatdetektiv Dixon Hill geladen.)

5.3 R2D2 und C3PO

„Er ist ein Protokoll-Droide und soll Mom helfen.[...]“ - „Oh! Hallo! Ich bin C-3PO, Roboter-Mensch-Kontakter.“

*- Anakin Skywalker und C-3PO; [**?**]*

Das *Star Wars*-Universum wimmelt von Robotern, oder, wie sie dort heißen, Droiden. Unter diesen Droiden finden sich R2D2 und C3PO, wohl zwei der bekanntesten Roboter der Filmgeschichte, vielleicht nur übertroffen vom Tin Man aus Oz. Während die Hauptfunktionsweise von R2D2 ist, verbale Befehle in simple Arbeitsanweisungen zu übersetzen und auszuführen, so arbeitet 3PO auf einem anderen Niveau: C-3PO akzeptiert eine (nahezu) beliebige Menschen-, Alien- oder Maschinensprache und kann sie in

(nahezu) jede andere Sprache übersetzen. Er wurde dazu konzipiert, die Kommunikation zwischen künstlichen und natürlichen Intelligenzen zu vereinfachen.
Beide nehmen sprachliche Anweisungen entgegen. Während R2 diese Anweisungen lediglich auf Befehle parst, die er kennt, analysiert 3PO seine erhaltenen Anweisungen und überführt sie in eine andere Sprache. Somit erfolgt bei R2 eine Reaktion auf Anweisungen und bei 3PO eine Umwandlung von Anweisungen.
Dies war zwar nicht vor langer, langer Zeit denkbar - zumindest nicht in dieser Galaxis - dürfte jedoch mehr als denkbar sein für die nicht allzu ferne Zukunft.

6 Schlussfolgerung

evtl mit Ausblick zusammenziehen

- Fazit: Status Quo
- Heutige Möglichkeiten
- theoretische Möglichkeiten
- Ansätze für Weiterentwicklung

Literatur

[1] Ewan Klein; Edward Loper; Steven Bird. *Natural language processing with Python*. O'Reilly Media, Inc, 2009.

[2] James H. Martin; Daniel Jurafsky. *Speech and Language Processing*. Prentice Hall, 2009.

[3] Clarity Lab. Sirius, 2015.

[4] Iain Last; Axel Glaser; Stefan Scheit; Tarmo Ploom. Processing chains in system of systems. In *Mesoca 2014: 8th IEEE International Symposium on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems*, 2014.

[5] Peter Norvig; Stuart Russel. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2009.

[6] Charles Marsh; Shubhro Saha. Jasper project, 2014.

[7] Adam Nash; Frederik Gladhorn; Mathias Stieger; Patrick von Reth; Peter Grasch. Simon listens, 2015.