Spracherkennung, ein gelöstes Problem?

Inhalt

Einleitung

Auf Apples Website ist zu lesen: „Siri hat die Antwort auf deine Fragen“ (Apple, 2020), und spätestens nach dem Erscheinen der Spracherkennungssoftware im Jahr 2011 (Wikipedia, 2020) sollte klar sein, dass tatsächlich der Fall eintreten wird, dass Spracherkennungssysteme uns vollends verstehen können.

In der Folgenen Arbeit soll es darum gehen, ob der Punkt bereits gekommen ist, an dem wir Spracherkennung an sich bereits als gelöstes Problem abhaken können und uns gänzlich der Verarbeitung und den Antworten die ein System zurückgeben soll konzentrieren können, und das Erkennen in jeglicher Situation besser oder mindestens gleich gut wie mit einem Menschen funktioniert.

Wichtige Begriffe

Zu Beginn muss klar abgegränzt werden, was genau mit Spracherkennung gemeint ist. Siri oder vergleichbare Systeme wie Amazon Alexa oder Google Home sind weit mehr als nur Spracherkenner. Es handelt sich um Komplexe sogenannte Assistenzsysteme welche nachdem die Sprache erkannt wurde noch den Sinn des Gesagten also die Bedeutung interpretieren und eine darauf angepasste, möglichst informative Antwort formulieren. Darüber hinaus werden unter dem Überbegriff Assistenzsystemen meist auch Systeme geführt, welche die Gewohnheiten und Aktivitäten analysiert und so das Benutzererlebnis verbessern sollen.

Für Spracherkenner an sich gibt es weitere Begrifflichkeiten, welche nützlich zu wissen sind:

Da wäre einerseits der begriff Neuronale Netze, dessen Kenntnis ich an dieser Stelle voraussetze. Neuronale Netze werden bei Spracherkennern in irgendeiner Form immer benutzt. Sie müssen mit Audiodaten trainiert werden um anschließend Audio und Wort oder Satz verknüpfen zu können. Hier kommen Sprachkorpora zum Einsatz, welche ich in einem späteren Abschnitt genauer betrachten werde.

Die Word Error Rate kurz WER wird häufig zur Wertung von Spracherkennern benutzt. Sie gibt an wie viele Wörter, prozentual zur Gesamtmenge, falsch erkannt wurden. Die WER kann für einen einzelnen Satz, aber auch für eine Maschine im Durchschnitt angegeben werden. Man geht im allgemeinen bei einem Durchschnittsmenschen von 6.0 aus.

Definition des Problems

Es bleibt die Frage nach dem konkret zu lösenden Problem und die Frage danach, wann genau eine Lösung tatsächlich das Problem gelöst hat.

Von E.Therhardt wird das Problem und seine Natur mit anderen Problemen und deren Lösungen verglichen wie zum Beispieldem Problem, Muskelkraft durch eine Maschine zu ersetzt. Es gibt hierfür eindeutig eine Lösung, welche für einen Anwendungsfall gültig ist. Seine folgende Definition zum Problem der Spracherkennung lautet wie folgt: „Man baue ein Gerät, welches imstande ist, gesprochene Sprache in einer geeigneten symbolischen Form, beispielsweise schriftlich, wiederzugeben“.

Obwohl auch von E. Therhardt im Folgenden noch deutlich gemacht wird, dass Spracherkennung weit komplexer ist, als ein einfaches technisches Problem, möchte ich für die Lösung eine etwas andere Definition verwenden.

Ich möchte mich sehr stark am Menschen orientieren, das hat zwei Gründe. Erstens gilt allgemein, dass wir möglichst intelligente Maschinen möchten. Bei einem Spracherkenner kann man diese Intelligenz jedoch nur schwer Messen. Zwar gibt es messbare Größen, diese sind jedoch nicht immer genau genug oder aussagekräftig genug, dazu jedoch später mehr. Zweitens ist es auch für einen Menschen nicht qualifizierbar ob er genug Sprache versteht oder nicht. Die Frage für den Menschen wäre folgende: Verstehst du gut genug um dein Sprachkenntnis nie wieder zu verbessern? Diese Frage sollte alleine auf der Basis des Sprichwortes „Man lernt nie aus“, von keinem Menschen jemals mit ja beantwortet werden. Daher scheint auch hier ein Durschnittsmensch als Maßstab angemessen, welcher bei Fertigstellung auch Nutzer des Spracherkenners sein wird. Ein Spracherkenner welcher genau so viel wie ein durchschnittlicher erwachsener Mensch erkennt, sollte demnach ausgereift sein. Ich möchte für diese Arbeit also annehmen, dass die Spracherkennung als gelöstes Problem gilt, sobald sie auf jeglichen Gebieten mindestens so gut wie ein Mensch funktioniert.

State oft the art

Heutige Spracherkennungssysteme sind bereits sehr ausgereift und werden in vielen Bereichen mehr oder weniger sinnvoll genutzt. So gibt es einerseits Spracherkennung in industriellen Bereichen, welche hauptsächlich für Diktiergeräte eingesetzt werden, um Protokolle oder Berichte in Textform umzuwandeln. Andererseits gibt es seit einiger Zeit Spracherkennung im Kommerziellen Sektor, wozu einerseits die Assistenzsysteme zählen, andererseits aber auch die Spracherkennung bei z.B. Telefonhotlines, welche meist mehr oder minder gut funktionieren. Alle Systeme dienen dazu, weniger Eingaben händisch vornehmen zu müssen. Persönlich und fern von jeglichen Messungen oder konkreten Werten möchte ich die heutigen Assistenzsysteme als bereits sehr ausgereift bezeichnen. Es passiert sehr selten, das etwas nicht verstanden wird und mit sehr hoher Wahrscheinlichkeit wird auch eine passsende Antwort auf das Gefragte geliefert. Die Systeme sind bereits so weit, das Personen ohne Technisches Verständnis die Vorteile gut und zuverlässig nutzen können. Ich erachte das für sehr positiv.

Historische Entwicklungen

Als erstes Spracherkennungssystem bezeichnen die meisten das „Audrey System“ aus dem Jahr 1952, welches zwar nur einzelne Zahlen und diese mit definiert großem Abstand zueinander erkannt hat, jedoch für seine Zeit schon weit in die Zukunft gedacht hatte. Es wurde, sobald auf einen Sprecher eingestellt wurde, eine WER von 0,03 oder weniger erreicht was selbst im Vergleich zu heutigen Systemen sehr niedrig ist. Alles natürlich davon abgesehen, dass es einen sechs fuß hohen Relay Schrank benötigte und das Eingeben von Zahlen meist schneller mit der Tastatur von statten geht. (Sonix, 2020)

Eine Weiterentwicklung des Audrey Systems war die 1962 erschienenen IBM Shoebox, welche bereits 16 Wörter erkennen konnte. Das System stellte einen vollwertigen, sprachgesteuerten Taschenrechner dar und basierte rein auf analoger Technik. (Sonix, 2020)

1970 stieß die DARPA (Defense Advanced Research Projects Agency) eines der bis heute größten Projekte zum Thema Spracherkennung in Gang: Das SUR oder auch Speech understanding researche.

Innerhalb dieses Projekts entstand Carnegie Mellon’s “Harpy’ speech system“. Zwar war das erste Ziel nur ein Vergleich zweier bestehender Systeme, schlussendlich entstand jedoch ein Hybrid aus beiden (Lowerre, 1976) welcher Sphynx genannt wurde und deutlich besser als alle bisher existierenden Systeme funktionierte.

1980 wurde der bisher größte Sprung bezüglich der Anzahl der erkannten Wörter durch die sogenannten Hidden Markow Modelle gemacht. Es wurden bisher nur einige hunderte Wörter erkannt und durch die Wahrscheinlichkeitsanalyse mit den Markow Modellen nun mehrere Tausende.

Durch die um die Jahrtausendwende immer schneller und besser werdenden Prozessoren wurden trivialer weise auch die Spracherekennungssysteme besser. Es entstanden erste sogenannte Voice Tree Systeme, welche eine Automatisierung mit Spracherekennung für Telefonhotlines darstellen. Diese sind mit derselben Technik teilweise heutzutage noch im Einsatz, was ihren sehr geringen Funktionsumfang erklärt. So findet man auf manchen Websiten Zitate wie: „VAL paved the way for all the inaccurate voice-activated menus that would plague callers for the next 15 years and beyond." (PC World, 2020)

Im selben Zug wurde 2000 ein System entwickelt welches bereits natürlich gesprochene Sprache, das heißt ca. 100 Wörter pro Minute erkennen konnte. Es war zwar weiterhin ein Training der Maschine für ca. 45 Minuten notwendig und der Preis des Systems belief sich auf rund 700$, dennoch war ein enormer Schritt zu vernehmen. (PC World, 2020)

In den folgenden Jahren geriet die Spracherkennung etwas in den Hintergrund. Zwar waren in Systemen wie Mac OS X und Windows Vista Sprachbefehle eingebaut, jedoch waren diese kaum jemandem bekannt. Des weiteren war das benutzen einer „altmodischen" Tastatur weiterhin einfacher und schneller.

Nachdem Die Google Voice Serach App für das IPhone erschienen ist, erlebte der Jubel um die Spracherkennung aus zwei Gründen einen erneuten Aufschwung: Erstens waren Smartphones welche sich zu der Zeit ganz neu auf dem Markt befanden ein Idealer Anwendungsbereich für Spracherkennung und zweitens, war es Google mittlerweile möglich, die Aufgenommene Sprache in eine Cloud zu laden, dort zu bearbeiten und das Ergebnis anschließend an den Absender zurück zu senden. Und all das in einer Zeit, welche dem Nutzer kaum oder gar nicht auffallen konnte. Auf diese Art und Weise funktioniert ein Großteil der heutigen Systeme immer noch. Dennoch war Spracherkennung zu jenem Zeitpunkt keinesfalls gelöst. Sie war auf sehr wenige Befehle beschränkt und aufgrund der Zuverlässigkeit wurde die Manuelle Eingabe meist bevorzugt.

Korpora gesprochener Sprache

Als Sprachkorpus oder Korpora gesprochener Sprache werden Sammlungen von Audiodaten bezeichnet, welche zum trainieren der Neuronalen Netze der Spracherkennungssysteme notwendig sind.

Aufbau und Nutzen

Sprachkorpora bestehen meist aus einer oder mehreren Audiodateien, für welche die Orthografische Transkription und die Zeitintervalle der Worte in einem Textfile festgehalten werden. Darüber hinaus enthalten manche Korpora eine Transkription in die Phonem Darstellung.

Allgemein lassen sich die meisten Sprachkorpora in zwei Gruppen einteilen: Einerseits gibt es die Korpora, welche auf gelesener Sprache basieren. Ein geschriebener Text wird hier von einer Person vorgelesen. Die Transkription existiert so bereits und es müssen nur noch die Zeiten herausgearbeitet werden. Es handelt sich oft um Hörbücher, Nachrichtensendungen oder auch Wortlisten. Zum anderen gibt es noch die Sprachkorpora welche auf „spontaner Sprache" basieren. Diese Sprachkorpora umfassen beispielsweise aufgenommene Dialoge, einfache Erzählungen, Wegbeschreibungen oder auch erzwungene Konversationen, bei welchen Probanden ein Gesprächsthema abarbeiten müssen.

Aktuelle Korpora

Heutzutage gibt es beinahe endlos viele Sprachkorpora, welche auch kommerziell nutzbar sind. Sie umfassen meist mehrere Stunden Sprache und verschiedene Sprecher. Es gibt Sprachkorpora in etlichen Sprachen, jedoch am meisten in Englisch. Weit hinterher sind die Sprachen der Ländern, welche nicht weit entwickelt sind, obwohl sie einen sehr großen Teil der Weltbevölkerung abdecken.

Auch das kann in Bezug auf die Lösung des Problems betrachtet werden: Zwar ist Spracherkennung für einige Sprachen aufgrund der Datenbasis bereits sehr ausgereift, in anderen Sprachen fehlt jedoch eine Datenbasis gänzlich. Es müssten etliche Korpora erstellt werden um zum Beispiel einen Indischen Dialekt verstehen zu können. Bisher gibt es leider keine Lösung für den enormen Hunger an Daten der Spracherkenner. In Bezug auf diese Eigenschaft wird es also noch sehr lange dauern bis das Problem Spracherkennung für alle Sprachen gelöst ist.

Der Mensch als Sprachkenner

Nach Noam Chomsky ist „Sprache das wichtigste Medium der zwischenmenschlichen Kommunikation, gleichzeitig aber auch eine der kompliziertesten Fähigkeiten des Menschen.“ Chomsky nach (Günther, 2007).

Dabei erschließt sich der Mensch Sprache aus Wortschatz und Grammatik was jedoch nicht den vollen Umfang von Sprache an sich umfasst, an dieser Stelle jedoch genügt. Der Mensch lernt ab dem ersten Tag mit Sprache umzugehen. Bis zum zwölften Monat hat ein Kind ca. 120 Stunden Sprache, was zwischen 500 000 – 1 000 000 Wörter sind, gehört und ein Erwachsener nimmt pro Jahr im Schnitt 2 000 Stunden Sprache auf, was rund 14 000 000 Wörter entspricht. Im Vergleich zur Maschine hat der Mensch also eine viel größere Datenbasis. Diese ist zwar nicht so detailiert und genau, das heißt es hat nicht jedes Wort eine schriftliche Bedeutung, sondern die Datenbasis des Menschen ist sehr vielfältig. Kein Kind hat das Wort Mama gelernt ohne dabei seine Mutter vor Augen zu haben. Der Mensch lernt also mit all seinen Sinnen, was ein Vorteil gegenüber der Maschine ist, welche nur auditiv arbeiten kann. Daher können Menschen auch Worte verstehen, welche sie eigentlich nicht verstanden haben. Was zuerst irrational klingt wird an einem Beispiel sehr schnell deutlich: „Gestern ging ich zum B… um Brötchen zu kaufen“. Jedem wird sofort klar, der Vollständige Satz das Wort Bäcker enthält. Wir können somit Wörter aus dem Kontext erschließen, was unser Gehirn „von ganz alleine“ macht.

Dieses Kontextwissen aus Visuellen und anderen Eindrücken und aus Erfahrung hat eine Maschine nicht. Um so gut wie ein Mensch zu werden muss die Maschine diese Schwäche an anderer Stelle Ausgleichen. Eine Maschine muss so Wörter viel besser verstehen um insgesamt ein annährend so gutes Ergebnis wie ein Mensch zu erhalten. Es reicht also rein auf der Wörter erkennen Ebene nicht, nur so gut wie der Mensch zu sein, da der Mensch zusätzlich zu den Worten die er erkennt, Dinge erschließen kann.

Dieser Teil sollte bei einer Maschine von einem anderen System erschlossen werden. Hierzu möchte ich später mehr erklären

Der Unterschied zur Maschine

Maschinen sind also nicht oder nur in begrenztem Maße fähig, Dinge aus dem Kontext zu schließen. Es gibt für eine Maschine genau zwei Möglichkeiten: Das Wort wird richtig verstanden oder das Wort wird nicht richtig verstanden. Fehlt einer Maschine also wie bei dem oben genannten Beispiel das Wort „Bäcker“, so geht ein erheblicher Teil der Information verloren. Der reine Spracherkenner, das heißt ohne Deutung der Worte in einen Sinn ist hier also bei weitem nicht so mächtig wie der Mensch.

Zukunftsausblick, was sagen die Zahlen?

Bereits 2016 wurde von einem Team bei Microsoft gemeldet, dass die Marke von 6.0 WER unterboten wurde. Microsoft schreibt „The researchers reported a word error rate (WER) of 5.9 percent, down from the 6.3 percent WER the team reported just last month.“ (Microsoft, 2020). Was ist bei einer so rasanten Entwicklung in Zukunft zu erwarten?

Trotz dass die WER nun kleiner als die des Menschen ist wird sie vermutlich in den nächsten Jahren weiter schrumpfen. Es sind zum aktuellen Zeitpunkt zwar keine neueren Daten bekannt, dennoch gehe ich stark davon aus, dass bereits Maschinen mit einer WER von kleiner als 5,0 existieren. Sie sind auf dem Papier also besser als der Mensch. Des weiteren ist es für die Maschinen theoretisch möglich eine WER von nahe 0 zu erreichen. Wann das jedoch der Fall sein wird, kann man nicht genau sagen. Was aber mit Sicherheit gesagt werden kann ist, dass eine Maschine mit einer WER von nahe Null das Problem vermutlich gelöst hat. Mit Sicherheit kann man das jedoch erst behaupten wenn man den direkten Vergleich mit einem Menschen eingeht und betrachtet wie sich beide in allen möglichen Varianten schlagen. Doch es existiert ein weiteres Problem:

Probleme der Word Error Rate

Die Word Error Rate nimmt keinerlei Rücksicht darauf, welches konkrete Wort in einem Satz verstanden wird und welches nicht. So wird zum Beispiel aus dem Satz: „Was gibt eins plus eins“ durch das nicht verstehen des Wortes plus eine nicht verwertbare Information. Die WER würde sich hierbei auf 16,6 belaufen. Gehen wir jedoch davon aus, dass „Was gibt“ nicht verstanden wird, haben wir eine WER von 33,3, also wesentlich höher, der Sinn bzw. die Information des Satzes bleibt aber erhalten. Das Ergebnis würde mit einer großen Wahrscheinlichkeit richtig ausgegeben werden.

Die Word Error Rate als Maßstab zu nehmen ist also sehr trügerisch. Zwar macht sie für den Vergleich für Maschinen untereinander Sinn, jedoch ist ein vergleich mit dem Menschen auf Basis der WER nicht sinnvoll. Darüber hinaus wird dieses Argument noch durch das Kontextwissen des Menschen, welches ich im Abschnitt „Der Unterschied zur Maschine“ angesprochen habe, verstärkt. Es ist als Mensch möglich, sich trotz eines fehlenden Wortes, den ganzen Satz zu erschließen, wodurch die WER, in einem einzelnen Satz, von z.B. 16,6 auf 0,0 herabsinkt.

Verbesserungspotential

Besser, jedoch auch noch nicht perfekt im Hinblick auf das Kontextwissen des Menschen, ist die Key Word Error Rate, kurz KER. Hierbei werden Worte, welche den Sinn des Satzes maßgeblich bestimmen, wie zum Beispiel Substantive und Verben, höher gewichtet als „Füllwörter“. So kann zum Beispiel je nach Kontext eine Wortgruppe stärker gewichtet werden als eine Andere. Erreicht wird dadurch, dass der Wert KER aussagekräftiger im Hinblick auf das Verstehen der Sprache ist. Dieses Vorgehen beschreibt jedoch noch lange nicht das Kontextwissen und die Erfahrung des Menschen.

Resümee

Zu Beginn habe ich darüber gesprochen, dass uns in diesem Artikel hauptschlich die Spracherkennung und nicht das Verstehen interessiert. Der Spracherkenner als solches wird meistens an der WER gemessen. Wir haben von 19…….. an gesehen, das eine stetige Entwicklung von nur wenigen Worten bis hin zu tausenden Worten erfolgreich gemeistert wurde. Heutzutage befinden wir uns in der Situation, dass die WER bereits besser als die des Menschen ist und auch weiterhin

immer besser werden wird. Doch ist das eine Lösung des Problems. Diese Frage ist nur teilweise mit Ja zu beantworten. Ich möchte mich zuerst auf diese Bereiche beziehen:

Betrachten wir nur Sprachen, für welche es etliche Sprachkorpora von verschiedenster Herkunft und insgesamt eine große Datenbasis gibt, ist es möglich einen Spracherkenner zu bauen, welcher in einem durchschnittlichen Satz mindestens gleich viel oder mehr Wörter als der Mensch erkennt. Für den Fall das die Systeme, welche das Erkannte verarbeiten, gleich gut wie ein Mensch arbeiten, haben wir also einen Spracherkenner, welcher gleich gut oder Besser ist als der Mensch.

Anhand dieses Abschnittes möchte ich endgültig zeigen, weshalb das Problem keinesfalls als gelöst betrachtet werden darf:

Damit das Problem als gelöst betrachtet werden kann, muss von einer einzelnen Sprache mit einer großen Datenbasis ausgegangen werden. Das trifft nur auf verhältnismäßig wenig Sprachen zu. Das Problem substituiert sich hier auf das Problem des Datenmangels. Da bisher keine andere Möglichkeit als immer mehr Daten zur Verbesserung der Spracherkenner gefunden worden ist, muss zuerst dieses Problem gelöst werden, um das Problem der Spracherkennung zu lösen.

Sollte dieses gelöst sein und es kommt dazu, dass jede Sprache gleich erkannt wird, sind wpr der Lösung einen Schritt näher. Nehmen wir gleichzeitig an, dass die WER bis dahin ebenfalls nahe null gesunken ist, so scheint der Spracherkenner an sich als gelöst.

An dieser Stelle muss nun jedoch trotz meiner anfänglichen Aussage, hierfür nur den Spracherkenner und nicht die Verarbeitung zu betrachten, das nachgeschaltete System betrachtet werden, welches bisher noch sehr unterentwickelt ist. Zwar ist die reine Spracherkennung gelöst, jedoch reicht die Intelligenz des Systems meiner Meinung nach nicht aus um dem Menschen ebenbürtig auf dem Gebiet Sprache erkennen zu sein. Es ist ein Neuronales netz notwendig, welches nicht nur einzelne Worte oder Wortgruppen erkennt, sondern Sätze mit Lücken anhand des Kontextes füllen kann. Der Kontext welchen ich bisher undefiniert benutzt habe umfasst aber weit mehr als nur die Worte. Stimmlage, Vorausgegangenes Gespräch, Ort und auch Zeitpunkt sind ausschlaggebend.

Auch wenn es so scheint als wäre der größte Teil der Spracherkennung bereits gelöst, so haben wir den komplexesten Teil noch vor uns.

An dieser Stelle möchte ich noch einen kleinen Zukunftsausblick geben. Der Sprachassistent von Google kann nun schon seit einigen Jahren auf Folgefragen eingehen. Was bereits einen teil des Kontextes mit einbezieht. Ich bin aufgrund dessen zuversichtlich, dass das Problem Spracherkenner in einer unbestimmten Zukunft kein Problem mehr sein wird.

Literaturverzeichnis

Apple. (2020). *Apple*. Von https://www.apple.com/de/siri/ abgerufen

Günther. (2007).

Günther, C. n. (2007). Seite 31.

Lowerre, B. T. (April 1976). *The HARPY Speech Recognition System.* Von https://stacks.stanford.edu/file/druid:rq916rn6924/rq916rn6924.pdf abgerufen

Microsoft. (September 2020). Von https://blogs.microsoft.com/ai/historic-achievement-microsoft-researchers-reach-human-parity-conversational-speech-recognition/ abgerufen

PC World. (August 2020). Von https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html?page=2 abgerufen

Sonix. (2020). *Sonix*. Von https://sonix.ai/history-of-speech-recognition abgerufen

Wikipedia. (30. August 2020). *Wikipedia*. Von https://de.wikipedia.org/wiki/Siri_(Software) abgerufen