Brandenburgische Technische Universität Cottbus – Senftenberg

Fakultät Maschinenbau, Elektrotechnik und Wirtschaftsingenieurwesen

Lehrstuhl Kommunikationstechnik

Prof. Dr.-Ing. habil. Matthias Wolff

Themenbetreuung: Dr.-Ing. Ronald Römer

Studienarbeit

Nachbildung der Hilbergschen Sprachmaschine

Michael Pöschick

Wallstraße 4

03185 Peitz

2707377

Michael.Poeschick@yahoo.de

0173/3868782

Abgabedatum: 23.05.2015

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

1.1. Stand der Technik und Problemstellung

Hirnforscher haben bis heute nicht herausgefunden, wie der Mensch denkt und wie Gedanken entstehen. (Vgl. [Hil12], S.11) Sicher scheint nur, dass die ca. 100 Milliarden Neuronen, welche sich kaum voneinander unterscheiden, etwas damit zu tun haben. Jedes dieser Neuronen besitzt zudem im Durchschnitt ca. 10000 Verbindungen zu anderen Neuronen (Vgl. [Hil12], S.14).

Viele Wissenschaftler orientieren sich bei der Suche nach dem Mechanismus des Denkens am Computer. Dadurch, dass Computer mit Hilfe von Symbolen arbeiten, sind sie aber grundsätzlich gar nicht zum sprachlichen Denken geeignet. Denn man ist sich mittlerweile sicher, dass Neuronen keine Symbole speichern. Vielmehr repräsentieren sie ihren Inhalt ausschließlich durch ihr Vorhandensein an einer ganz bestimmten Position im Neuronen-Netzwerk. (Vgl. [Hil12], S.13-14)

Diese Beobachtungen bilden die Grundlage für die „Sprachmaschine“, ein Projekt von Professor Wolfgang Hilberg an der TU Darmstadt.

Dessen Ziel ist es, Netzwerke nach dem Vorbild des menschlichen Gehirns aufzubauen, sodass sie zu höheren sprachlichen Leistungen fähig sind. Dazu sind heutige Computer und im Speziellen die Künstliche Intelligenz nicht zufriedenstellend in der Lage. (Vgl. [Hil12], S.11)

Im Speziellen ist das Ziel dieser Arbeit, mit Hilfe von Netzwerken ein hierarchisches Kodier-/ Dekodierverfahren zu realisieren, welches zu deutlich höheren Kompressionsraten fähig ist als nicht-hierarchische Systeme. Dies wird durch stufenweise Abstraktion von Texten erreicht, wobei am Ende nur noch die wichtigsten Informationen, vergleichbar mit den Kerngedanken, enthalten sind.

1.2. Der Ansatz der Sprachmaschine

Im Zuge der Entwicklung einer Sprachmaschine nimmt man an, dass jedes Wort, welches ein Mensch kennt, von einem Neuron in seinem Gehirn repräsentiert wird. Es gibt allerdings sehr viel mehr Neuronen als Wörter, weswegen angenommen wird, dass die anderen Neuronen Sätze, komplette Texte oder Gedanken beinhalten könnten. (Vgl. [Hil12], S.12)

Der Wortschatz der deutschen Sprache umfasst gegenwärtig 300000 bis 500000 Wörter. (Vgl. [Dud09]) Setzt man davon nun zufällig ausgewählte Wörter hintereinander um einen Text zu bilden, gibt es davon unzählige Varianten. Schon für die Wahl von nur 3 Wörtern aus einem Wortschatz von 300000, gibt es $2{,}7\ x\ 10^{16}$ verschiedene Möglichkeiten. Um einen sinnvollen Text zu bilden sind deutlich mehr Wörter nötig, wodurch die Anzahl der möglichen Kombinationen ins Unermessliche ansteigen würde.

In einem sprachlich korrekten deutschen Text kann allerdings auf ein Wort nicht jedes beliebige andere Wort folgen. Auf das Wort „Fisch“ kann z.B. nicht das Wort „Staubsauger“ folgen. Lässt man nun für die zufällige Auswahl eines Wortes nur diejenigen Wörter zu, die auf das vorhergehende folgen können, sinkt die Anzahl der möglichen Wortkombinationen und damit der möglichen Texte erheblich.

Darauf aufbauend ist es möglich, Wörter aus Texten zu löschen und diese oder andere passende Wörter später wieder hinzuzufügen. Die Sprachmaschine nutzt das um Texte zu kodieren und zu komprimieren.

Mit Hilfe von großen Textsammlungen wird dafür eine Datenbank erstellt, welche die in den Texten benutzte Sprache darstellt. Darin sind im (unrealistischen) Optimalfall alle Wörter der Sprache enthalten, sowie Informationen darüber, welche Wörter auf bestimmte Wörter folgen können und welche Wörter mit dem Abstand zwei, drei, usw. folgen können.

Dies geschieht auf mehreren Ebenen. In der untersten, der Basisebene, sind alle in den Lerntexten enthaltenen Wörter vorhanden. Um von einer Ebene auf die nächsthöhere Ebene zu gelangen und damit die Information zu verdichten, werden in der unteren Ebene diejenigen Wörter gelöscht, die im Allgemeinen am häufigsten Vorkommen. Dies wird im einfachsten Fall

dadurch realisiert, dass von zwei aufeinander folgenden Worten jeweils das unwichtigere, häufigere gelöscht wird. Dieser Vorgang wird Abstraktion genannt. (Vgl. [Hil12], S.17)

Das seltenere, informationsreichere Wort wird in die darüber liegende Ebene übertragen und repräsentiert dort nicht nur genau das Wortpaar, welches davor aufgelöst wurde, sondern alle möglichen Wortpaare, die genau dieses Wort enthalten und in den Lerntexten mindestens einmal vorgekommen sind. Dieses Metawort bildet nun ein Wortpaar mit einem anderen Metawort dieser Ebene und wiederum wird das häufigere gelöscht und das seltenere in die nächsthöhere Ebene übertragen. (Vgl. [Hil12], S.18)

Ein Wort in der dritten Ebene repräsentiert dann 4 Wörter der untersten Ebene. Die Wörter der höheren Ebenen, die in der Basisebene z.B. einen kompletten Satz repräsentieren, kann man mit einem Gedanken vergleichen.

Im Gegensatz zur Künstlichen Intelligenz ist die Sprachmaschine nicht auf äußerst komplexe Grammatikregeln angewiesen, da diese in den Netzwerkstrukturen implizit enthalten sind. Im Optimalfall sind erzeugte Texte auch ohne Anwendung von Regeln grammatikalisch korrekt. (Vgl. [Hil12], S.19)

Im Rahmen dieser Arbeit werden diese Netzwerkknoten zwar mit Hilfe von Kodes simuliert, man könnte es aber mit entsprechendem Aufwand auch ohne Kodes realisieren.

Der Umfang dieser Arbeit erlaubt nur die Programmierung und Beurteilung der grundlegenden Algorithmen und Programmteile. Für eine in der Praxis einsetzbare Sprachmaschine sind zusätzliche Erweiterungen nötig.

2. Informationstheoretische Grundlagen

2.1. Information

Information ist ein von einem Sender (Quelle) über ein bestimmtes Medium (Kanal) an einen Empfänger (Senke) gesendetes Wissen. (siehe Abbildung 1)

Abbildung 1: Einfachstes Schema eines Nachrichtenübertragungssystems (Quelle: Eigene Darstellung in Anlehnung an [Göb07], S. 17)

Eine Information kann nur vorliegen, wenn über die Bedeutung eines Symbols, einer Geste, einer Form, oder Ähnlichem eine Vereinbarung getroffen wurde. So kann die Abkürzung „DFB“ für den einen Menschen, der die Bedeutung dieser Abkürzung kennt, eine Information darstellen und für einen anderen nicht. Demzufolge kann diese Information auch verloren gehen, obwohl die Symbole noch vorhanden sind, wenn diese Vereinbarung keinen Bestand mehr hat oder vergessen wurde, wie z.B. bei antiken Inschriften. (Vgl. [Hil12], S. 48-51) Hat man die russische Sprache nie erlernt, stellt ein russischer Text keine verwertbare Informationsquelle dar.

Der Informationsgehalt eines Symbols ist dabei stets proportional zum Grad der „Überraschung“. Ein Wort, das eher selten vorkommt, überrascht uns mehr als ein Wort, welches häufig vorkommt. Dieses Wort hat damit einen höheren Informationsgehalt. Für das eigentliche Übermitteln von Informationen sind in einem Text vor allem die seltenen Wörter von Bedeutung. Die häufigen Wörter dienen z.B. der Lesbarkeit.

2.2. Entropie und Redundanz

Die Entropie entspricht dem mittleren Informationsgehalt einer Quelle und wird auch als Informationsdichte bezeichnet (Vgl. [Sch12], S. 16):

$$H_m = -\sum_{i=1}^{N} p(x_i)\ ld\ p(x_i) \tag{1}$$

$Wobei\ p(x_i)$ die Auftretenswahrscheinlichkeit des Elements x_i ist und $-ld\ p(x_i)$ dessen Informationsgehalt. Je niedriger die Auftretenswahrscheinlichkeit eines Elements ist, umso größer ist dessen Informationsgehalt. Am größten ist der mittlere Informationsgehalt, wenn alle Symbole gleichwahrscheinlich auftreten. Die Entropie hat als die Maßeinheit z.B. bit/Zeichen. (Vgl. [Göb07], S. 21-25)

Redundanz beschreibt den Teil einer Nachricht, der keine Informationen enthält. Entfernt man die redundanten Stellen, bleibt der Informationsgehalt der Nachricht erhalten. Man erhöht dadurch die durchschnittlich übertragene Information pro Zeichen. Redundanz kann aber auch dazu genutzt werden die Störanfälligkeit bei Übertragungen zu reduzieren, wodurch die Qualität der Nachricht auf Kosten der zu übertragenden Datenmenge erhöht wird. (Vgl. [Sch12], S. 41)

Durch das schrittweise Entfernen von Redundanz beim Kodieren mit der Sprachmaschine wird die Entropie des kodierten Textes erhöht und ermöglicht dadurch enorme Kompressionsraten.

2.3. Quellenkodierung/ Optimalkodierung

Abbildung 2 zeigt die Informationsübertragung mit Hilfe von Kodes. Quellenkodierung beschreibt Verfahren, die den Zweck haben, Nachrichten auf ihren relevanten und nicht-redundanten Teil zu reduzieren. Dadurch sollen verfügbare Übertragungskanäle effizient genutzt werden. (Vgl. [Göb07], S. 28) Im Gegensatz dazu steht die Kanalkodierung, bei der

durch Hinzufügen von Redundanz die Störungsanfälligkeit bei der Übertragung reduziert werden soll. (Vgl. [Sch12], S.41)

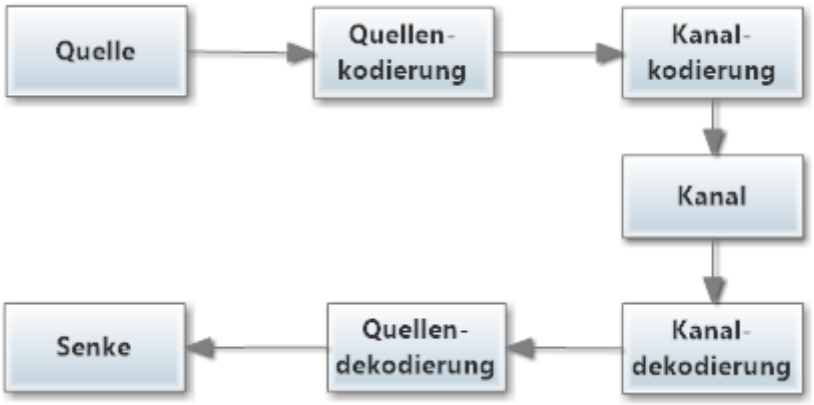

Abbildung 2: Schema eines Nachrichtenübertragungssystems mit Kodierung (Quelle: Eigene Darstellung in Anlehnung an [Sch12], S. 41)

Vollkommen redundanzfreie Kodes sind möglich, allerdings wird darauf aufgrund von großem Zeitaufwand in der Praxis verzichtet. Optimalkodierung ist der Oberbegriff für Methoden, welche mit vertretbarem Aufwand redundanzarme Kodes erzeugen. (Vgl. [Sch12], S.48-49)

Unterteilen lassen sich die Methoden in diejenigen, die eine Quellenstatistik voraussetzen und diejenigen, die dies nicht tun. Eine Quellenstatistik gibt Auskunft über die Auftrittswahrscheinlichkeit jedes verwendeten Zeichens. Je häufiger ein Zeichen auftritt, umso kürzer sollte dessen Kodewort sein. Seltener auftretende Zeichen bekommen dementsprechend ein längeres Kodewort. Dadurch wird die mittlere Kodelänge minimiert. Beispiel dafür sind das Shannon-Fano-Verfahren und das Huffman-Verfahren. (Vgl. [Sch12], S.48-49)

Beim Lempel-Ziv-Verfahren wird im Text nach wiederholt vorkommenden Zeichenketten gesucht, welche dann in einem Kodewort zusammengefasst und in einem dynamischen Wörterbuch gespeichert werden. (Vgl. [Sch12], S.59-60)

Die Methode bei der Sprachmaschine ist eine Mischung dieser Beiden. Es wird eine Quellenstatistik erstellt. Allerdings bekommen die häufigeren Zeichen bzw. Wörter keine kürzeren Kodewörter, da sie im kodierten, zu übertragenden Text in der Regel gar nicht mehr

enthalten sind. Das Zusammenfassen der wiederholt vorkommenden Zeichenketten ist mit dem Erzeugen der Metawörter vergleichbar.

3. Grundlagen für die Sprachmaschine

3.1. Hierarchieebenen

Das Kernprinzip der Sprachmaschine sind die Hierarchieebenen. Abbildung 3 zeigt schematisch die ersten 5 Ebenen eines Textes. Die erste, unterste Ebene ist die Ebene der Klarwörter und enthält den kompletten Fließtext. Jedes Quadrat entspricht hier genau einem Wort. Die zweite Ebene enthält dann anstelle von 2 Klarwörtern der ersten Ebene nur noch ein sogenanntes Metawort. Dessen Bezeichnung kann völlig unabhängig von den bildenden Wörtern sein. Um die Metawörter der höheren Ebenen aber besser nachvollziehen zu können, entsprechen sie jeweils dem selteneren der beiden Wörter. Diese Metaworte repräsentieren alle möglichen Kombinationen aus zwei Wörtern, die dieses seltenere Wort enthalten. Das häufigere Wort, welches hier der Redundanz entspricht, wurde entfernt.

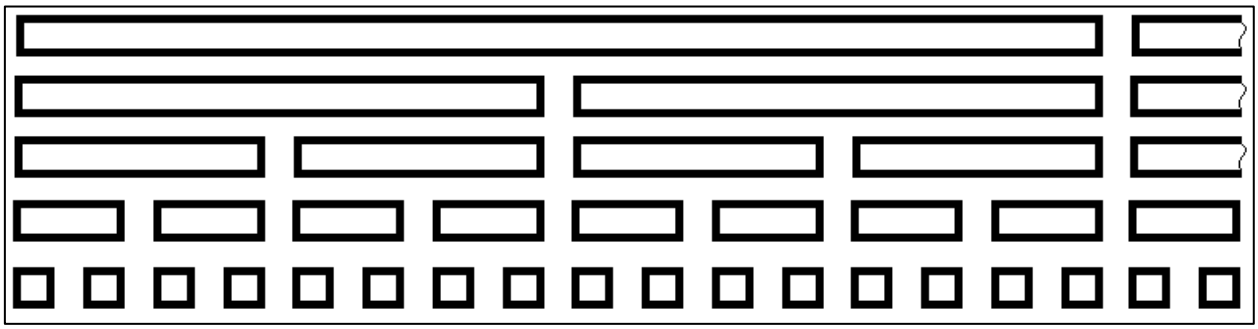

Abbildung 3: Schematische Darstellung der Hierarchieebenen (Quelle: Eigene Darstellung)

Dieses Prinzip wird entsprechend in den höheren Ebenen fortgesetzt, wobei dann nicht mehr zwei Klarwörter, sondern zwei Metawörter zu einem neuen Metawort in der nächsten Ebene werden. Ein Metawort in der 5. Ebene repräsentiert dann bereits 16 Wörter in der untersten Ebene. Anstelle dieser 16 Wörter müsste beim Kodieren mit 5 Ebenen also nur ein einziges Wort übermittelt werden.

3.2. Verbindungsmatrix

Um Wörter und ihre möglichen Nachfolger grafisch darzustellen, kann man einen gerichteten Graphen mit den Wörtern als Knoten benutzen. Da diese aber ab einer gewissen Anzahl von Wörtern unübersichtlich werden, nutzt man Verbindungsmatrizen.

Abbildung 4 zeigt einen Ausschnitt aus einer Verbindungsmatrix. Auf beiden Achsen stehen jeweils alle Wörter, absteigend geordnet nach ihrem Rang. Der Rang ergibt sich dabei entweder aus der absoluten Häufigkeit des Wortes, oder aus dem hier verwendeten Verzweigungsgrad eines Wortes, also der Anzahl der möglichen Nachfolger. Eine „1“ im Feld 4(links):3(oben) bedeutet, dass das Wort „die“ Nachfolger von „und“ sein kann. Gleichzeitig bedeutet es auch, dass „und“ Vorgänger von „die“ sein kann.

	1	2	3	4	5	6	7	8	9	10
1	[]	'der'	'die'	'und'	','	'.'	'den'	'"'	'von'	'des'
2	'der'	1	1	[]	[]	[]	1	1	1	[]
3	'die'	1	1	[]	1	[]	1	1	1	1
4	'und'	1	1	[]	1	[]	1	1	1	1
5	','	1	1	1	[]	1	1	1	1	1
6	'.'	1	1	1	1	[]	1	1	1	1
7	'den'	1	1	[]	[]	[]	[]	1	1	[]
8	'"'	1	1	1	1	1	1	1	1	1
9	'von'	1	[]	1	[]	[]	1	1	[]	[]
10	'des'	[]	[]	[]	[]	[]	[]	1	1	[]

Abbildung 4: Ausschnitt einer Verbindungsmatrix (Quelle: Eigene Darstellung)

Neben Verbindungsmatrizen für direkte Nachfolger und Vorgänger gibt es auch Verbindungsmatrizen mit höheren Reichweiten. Wörter, die als zweites auf ein bestimmtes Wort folgen können, können dementsprechend der Verbindungsmatrix der Reichweite 2 entnommen werden.

Zum Erstellen der Verbindungs- und Reichweitenmatrizen werden Lerntexte benötigt, welche einen Querschnitt durch die verwendete Sprache darstellen.

Die zur Anwendung gekommenen Texte entstammen dem LIMAS-Korpus. Dieser umfasst insgesamt 500 Texte mit je ca. 2000 Wörtern. Die Texte sind in 33 Rubriken, darunter z.B. Kultur, Geschichte und Wirtschaft, mit diversen Unterrubriken unterteilt. Bei einer relativ

kleinen Auswahl an Texten sollten Texte mit ähnlicher Sprache und sinnverwandten Themen gewählt werden, um Assoziationen über die Grenzen eines Textes hinweg zu ermöglichen. Verarbeitet wurden deshalb 35 Texte der Unterrubrik Tagespolitik.

3.3. Prädiktion

Will man in einen teilweise vorhandenen Satz ein Wort einfügen, nutzt man dafür die Prädiktion. Liest man z.B. den unvollständigen Satz „Ich bin ___ dem Weg“, kommt man zu dem Ergebnis, dass möglicherweise das Wort „auf“ fehlt. Die Prädiktion nutzt die bereits bekannten Wörter um in den Verbindungsmatrizen der verschiedenen Reichweiten nach passenden Ergänzungen zu suchen. Abbildung 5 zeigt den schematischen Ablauf.

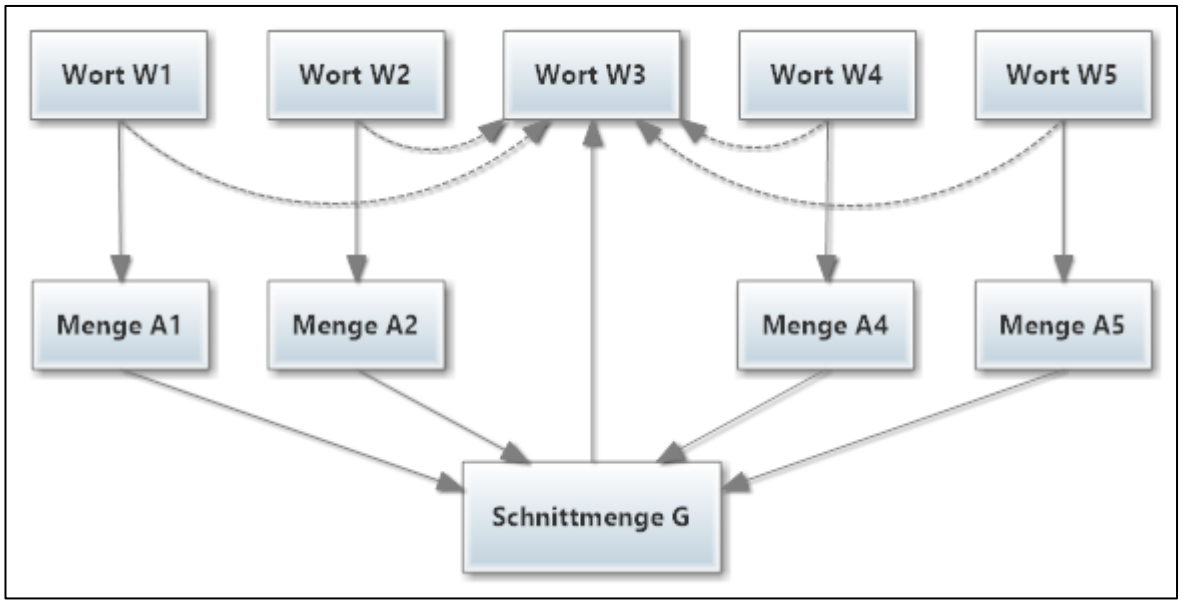

Abbildung 5: Schematischer Ablauf einer Prädiktion (Quelle: Eigene Darstellung in Anlehnung an ([Rie02], S.11)

Das Beispiel im Bild zeigt einen Fall, bei dem die Wörter W1, W2, W4 und W5 bekannt sind und ein ergänzendes Wort W3 gesucht wird. Mit Hilfe der Verbindungsmatrix der Reichweite 2 sucht man nun alle möglichen Wörter, die 2. Nachfolger von W1 sein können. Diese Wörter bilden die Menge A1. Mit der Verbindungsmatrix der Reichweite 1 sucht man alle Wörter, die direkter Nachfolger von W2 sein können. Diese bilden die Menge A2. Für die Mengen A4 und

A5 wird analog verfahren, nur dass anstatt den Nachfolgern die Vorgänger gesucht werden. Die Schnittmenge ergibt dann die Menge G und entspricht den Wörtern, die den Text sinnvoll ergänzen können.

Sollte Menge G aus mehreren Wörtern bestehen könnten zur weiteren Einschränkung Verbindungsmatrizen mit noch höheren Reichweiten eingesetzt werden. Sollte das nicht möglich oder nicht erwünscht sein, muss aus der Menge G ein Wort ausgewählt werden, z.B. das mit dem höchsten Rang.

3.4. Zipfsches Gesetz

Das Zipfsche Gesetz (Vgl. [Zip35]) wird vorrangig in der Linguistik eingesetzt und trifft Aussagen über die Häufigkeiten von unterschiedlichen Wörtern. Die Wörter eines Textes bzw. mehrerer Texte werden absteigend nach ihrer Häufigkeit sortiert und bekommen Ränge zugeordnet. Das häufigste Wort entspricht Rang 1, das zweithäufigste Wort Rang 2 usw.

Trägt man diese dann in ein Diagramm mit logarithmisch geteilten Achsen ein, wobei sich auf der x-Achse der Rang und auf der y-Achse die Häufigkeit befindet, ergibt sich immer eine abfallende Gerade (Vgl. [Rie02], S. 6). Anstelle der Häufigkeit wird hier der Verzweigungsgrad verwendet, die Wörter werden also nach der Anzahl der potenziellen Nachfolger sortiert. Auf die Reihenfolge hat dies keinen signifikanten Einfluss. (Vgl. [Hil05], S. 20)

Das Gesetz wurde von George Kingsley Zipf bereits in den 1930er Jahren entdeckt und seine Richtigkeit an Texten verschiedener Autoren und Sprachen festgestellt.

Darüber hinaus sagt das Gesetz aus, dass häufiger verwendete Wörter tendenziell kürzer sind.

Für die Sprachmaschine ist dieses Gesetz von Bedeutung, da es die optimale Grundstruktur einer menschlichen Sprache darstellt. Diese Grundstruktur ist dem Gehirn perfekt angepasst und muss deshalb auch in allen Abstraktionsebenen vorhanden sein. (Vgl. [Hil05], S. 21)

4. Systembeschreibung

Das im Kontext dieser Arbeit entstandene MATLAB-Programm dient dazu, die grundlegenden Algorithmen, die bei der Sprachmaschine eine Rolle spielen, auf ihre Funktionalität überprüfen zu können. Darüber hinaus soll das Zipfsche Gesetz bestätigt werden.

Wie Abbildung 6 zeigt, ist der erste Schritt das Umwandeln der ausgewählten Texte des LIMAS-Korpus in ein Format, mit dem das Programm arbeiten kann. Dabei werden unter anderem alle großen Buchstaben in kleine Buchstaben geändert und die im LIMAS-Korpus vorhandene Zeilennummerierung entfernt. Außerdem wird vor Sonderzeichen ein Leerzeichen eingefügt, damit diese als eigenständiges Wort betrachtet werden können.

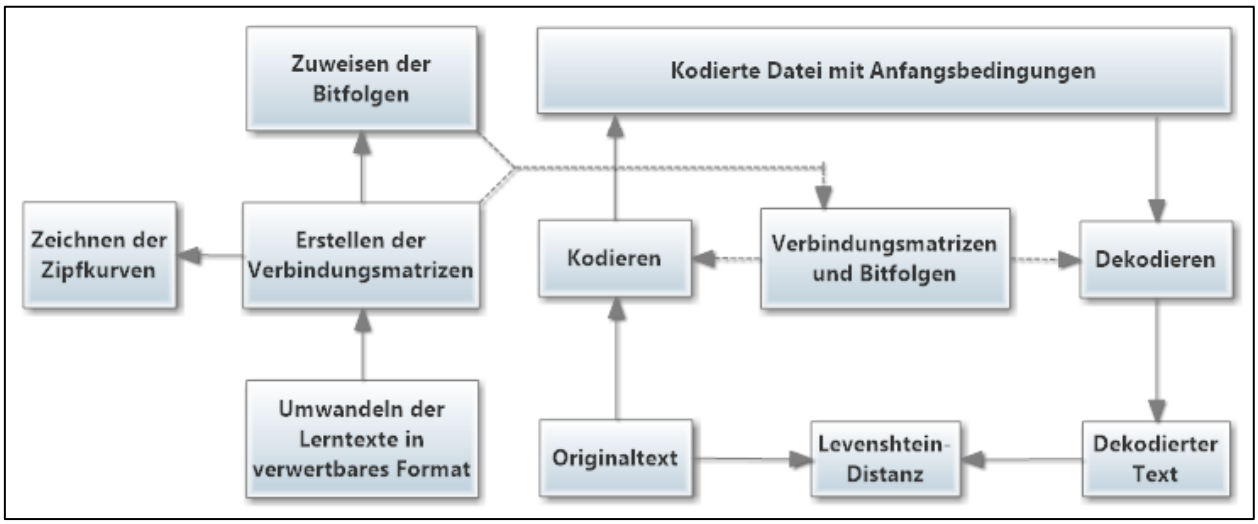

Abbildung 6: Ablaufdiagramm des MATLAB-Programms (Quelle: Eigene Darstellung)

Mit Hilfe der umgewandelten Texte werden die Verbindungsmatrizen der unterschiedlichen Reichweiten für die gewünschte Anzahl an Ebenen erstellt. Im Rahmen dieser Arbeit wird die Kodierung mit 5 Ebenen vorgenommen, weswegen Verbindungsmatrizen für die Ebenen 1 bis 4 erstellt werden, da die 5. Ebene im kodierten Text komplett vorhanden ist und deshalb keine Informationen über mögliche Zusammenhänge in dieser Ebene benötigt werden. Es werden Verbindungsmatrizen für die Reichweiten 1, 2 und 3 erstellt.

Anschließend können aus den Verbindungsmatrizen die dazugehörigen Zipfkurven erstellt werden. Dabei wird auf logarithmisch geteilten Achsen der Verzweigungsgrad der Wörter über deren Rang aufgetragen.

Außerdem wird jedem Wort eine 15-stellige Bitfolge zugeordnet. Das Wort „europa“ bekommt z.B. die Bitfolge 00000 01011 01110.

Die bisherigen Tätigkeiten dienten dem Aufbau der Datenbank bzw. dessen Analyse. Mit der geschaffenen Basis kann nun mit dem Kodieren und Dekodieren begonnen werden. Aus den 35 Texten, die die Grundlage bilden, werden 10 Texte ausgewählt, welche anschließend kodiert und dekodiert werden.

Die umgewandelten Originaltexte werden dann kodiert, indem von einem Wortpaar das häufiger vorkommende Wort, also das mit dem niedrigeren Informationsgehalt, gelöscht wird und das andere Wort in die nächste Ebene übergeben wird, wo dann genauso weiter verfahren wird. Für dieses Entfernen von redundanten Informationen werden die Ranginformationen aus den Verbindungsmatrizen genutzt. In der kodierten Datei befinden sich am Ende nur 4 Bits, welche die Anzahl der verwendeten Ebenen übergibt, die Bitfolgen der Wörter der letzten Ebene, sowie die der Startwörter jeder Ebene. Die Startwörter, die auch als Anfangsbedingungen bezeichnet werden, dienen später der Prädiktion als Einstieg.

Beim Dekodieren werden zunächst die Bitfolgen in die entsprechenden Wörter umgewandelt und anschließend wird mit Hilfe der Prädiktion nacheinander den Ebenen wieder Redundanz hinzugefügt, wobei mit Ebene 4 angefangen wird. Diese Redundanz entspricht dabei nicht zwangsläufig der beim kodieren entfernten Redundanz.

Um die Ähnlichkeit zwischen den Originaltexten und den dekodierten Texten festzustellen, wird als letzter Schritt jeweils die Levenshtein-Distanz berechnet.

5. Ergebnisse

5.1. Bestätigung des Zipfschen Gesetzes

Zur Bestätigung, dass das Zipfsche Gesetz für die 35 verwendeten Texte zutrifft, zeigt die Abbildung 77 die Zipfkurven für die Ebenen 1 und 2. Wie erwartet entsprechen die Graphen näherungsweise einer Geraden. Die Graphen für die Ebenen 3, 4 und 5 sehen ähnlich aus. Insgesamt sind 15461 verschiedene Wörter in den 35 Texten enthalten.

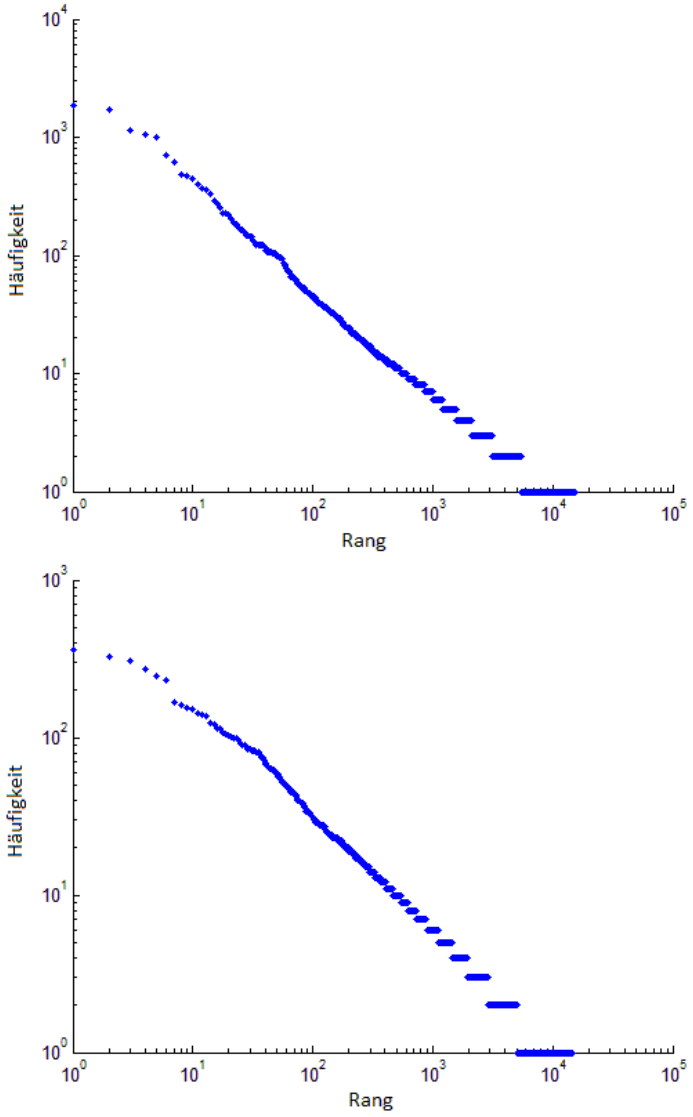

Abbildung 7: Zipfkurven für die Ebenen 1 und 2 (Quelle: Eigene Darstellung)

Das Zipfsche Gesetz besagt außerdem auch, dass kurze Wörter tendenziell häufiger verwendet werden. Tabelle 1 zeigt in der linken Spalte die 20 Wörter mit den meisten Verzweigungen innerhalb der untersuchten Texte. Die rechte Spalte enthält 20 der insgesamt 10050 Wörter, die nur einen Nachfolger haben. Diese Aussage konnte damit ebenfalls belegt werden.

Tabelle 1: Die 20 verzweigungsreichsten und -ärmsten Wörter der untersuchten Texte (Quelle: Eigene Darstellung)

der	naturgas
die	beigas
und	sorten
,	erdölprodukte
.	ausgangsstoffe
den	industriezweige
"	niedrigsiedenden
von	rohöldestillation
des	rohbenzin
das	leichtbenzin
in	hohem
zu	temperaturen
dem	gekrackt
eine	spaltprodukte
im	ausgangsstoff
ein	petrochemische
mit	synthesen
als	rohöl
nicht	schmierölen
einer	schmierfetten

5.2. Kodierungssergebnisse

In Zuge der Kodierung wird die Anzahl der Wörter eines Textes auf ca. $\frac{1}{16}$ der Originallänge gekürzt. Dieser Wert ist konstant für alle Texte der gleichen Länge und entsteht durch die Halbierung der Textlänge von Ebene zu Ebene und die verwendeten 5 Ebenen.

Dementsprechend würde die Länge des kodierten Textes bei Verwendung von 4 Ebenen ca. $\frac{1}{8}$ der Originallänge betragen und bei 6 Ebenen ca. $\frac{1}{32}$. Im Folgenden wird beispielhaft der LIMAS-Textes Nr. 21 verarbeitet.

Die ersten 100 Wörter (Sonderzeichen gelten als Wort) des Originaltextes:

Zusammengefaßte Ziele, Maßnahmen und Kosten. Zur besseren Übersicht werden die Ziele, Maßnahmen und Kosten des Nordrhein-Westfalen-Programms 1975 zunächst in der Kurzfassung der einzelnen Abschnitte dargestellt. Teil 3: Arbeit und Wirtschaft. Kernenergie Entwicklung von Hochtemperaturreaktoren mit unmittelbar angeschlossener Heliumgasturbine; größerer Anteil der Kernenergie an der Stromerzeugung; Anwendung von Prozeßwärme insbesondere zur Kohlevergasung, Erzverhüttung und Herstellung chemischer Rohstoffe. Bau eines 300-MWe-Hochtemperaturreaktors als Prototyp und Projektierung eines Leistungsreaktors mit Heliumgasturbine; Errichtung eines Sonderforschungsbereiches " Prozeßwärme " an der Kernforschungsanlage Jülich. Bergbauanpassung Steigerung der Leistung und Wettbewerbsfähigkeit des Steinkohlenbergbaues. Fortführung der

Die ersten 50 Wörter der 5. Ebene des Textes Nr. 21:

zusammengefaßte abschnitte heliumgasturbine chemischer leistungsreaktors bergbauanpassung bergbaues anteils kohletechnik bergmanns veredelung energiepreissenkung stromerzeugungsunternehmen stromerzeugungsunternehmen auflockerung problemgebieten bürgschaften produktions betriebseinheiten investitionsbeihilfen landwirtschaftlicher folgemaßnahmen dörfer wirtschaftswege altgehöften bearbeitungsstufe aufklräungsaktionen zukunftsaussichten verbesserte umschulungseinrichtungen älterer landesplanung abgedeckt landesentwicklungsplan gebietsentwicklungspläne stadtzentren bahnnetzes s-bahnhaltestellen 6ändlichen vorrangige mittelfristigen ballungskerne investitionsmittel inhalt programme ballungskerne ausgewählten preiswerten 250000 s-bahnhaltestellen

In der 5. Ebene sind wie erwartet nur sehr wenige kurze Wörter vorhanden. Lange, seltene und damit informationsreiche Wörter sind die Regel.

5.3. Dekodierungsergebnisse

5.3.1. Textbeispiele

Im Folgenden stehen jeweils die ersten 100 Wörter des vorverarbeiteten Originaltexts Nr. 21, sowie die mit unterschiedlichen Reichweiten dekodierten Ergebnistexte. Markiert sind dabei die Wörter, welche in Ebene 5 kodiert wurden (siehe vorheriges Kapitel). Alle anderen Wörter wurde mit Hilfe der Prädiktion hinzugefügt.

Original Nr.21

zusammengefaßte ziele , maßnahmen und kosten . zur besseren übersicht werden die ziele , maßnahmen und kosten des nordrhein-westfalen-programms 1975 zunächst in der kurzfassung der einzelnen abschnitte dargestellt . teil 3 : arbeit und wirtschaft . kernenergie entwicklung von hochtemperaturreaktoren mit unmittelbar angeschlossener heliumgasturbine ; größerer anteil der kernenergie an der stromerzeugung ; anwendung von prozeßwärme insbesondere zur kohlevergasung , erzverhüttung und herstellung chemischer rohstoffe . bau eines 300-mwe-hochtemperaturreaktors als prototyp und projektierung eines leistungsreaktors mit heliumgasturbine ; errichtung eines sonderforschungsbereiches " prozeßwärme " an der kernforschungsanlage jülich . bergbauanpassung steigerung der leistung und wettbewerbsfähigkeit des steinkohlenbergbaues . fortführung der

Reichweite 1

zusammengefaßte ziele in maßnahmen werden kosten der unternehmenskonzentration kläranlagen anzuschließen einen besseren übersicht werden andere ziele des nordrhein-westfalen-programms 1975 zunächst der kurzfassung der einzelnen abschnitte dargestellt . teil . 3 .märz 1970, mit sozialpartnern und wirtschaft . kernenergie von hochtemperaturreaktoren mit unmittelbar angeschlossener heliumgasturbine ; errichtung eines sonderforschungsbereiches die zunehmende stromerzeugung durch die anwendung von prozeßwärme insbesondere sibirien zur kohlevergasung , erzverhüttung herstellung chemischer rohstoffe wurden im bau eines 300-mwe-hochtemperaturreaktors als prototyp und projektierung eines leistungsreaktors mit

heliumgasturbine ; größerer 1975 errichtung eines sonderforschungsbereiches " an der kernforschungsanlage . bergbauanpassung steigerung des pädagogischen leistung und wettbewerbsfähigkeit des steinkohlenbergbaues . fortführung

Reichweite 2

zusammengefaßte ziele , maßnahmen werden kosten . zur besseren übersicht werden die ziele durch , maßnahmen werden kosten durch nordrhein-westfalen-programms gewährleisteten zunächst kurzfassung der einzelnen abschnitte dargestellt abschnitte teil . kritik 3 . arbeit und wirtschaft . kernenergie von hochtemperaturreaktoren mit unmittelbar angeschlossener heliumgasturbine ; größerer anteil der kernenergie an der stromerzeugung ; anwendung von prozeßwärme insbesondere zur kohlevergasung , erzverhüttung und herstellung chemischer rohstoffe . bau eines 300-mwe-hochtemperaturreaktors als prototyp und projektierung eines leistungsreaktors mit heliumgasturbine ; errichtung eines sonderforschungsbereiches " prozeßwärme " an der kernforschungsanlage jülich . bergbauanpassung steigerung der leistung und wettbewerbsfähigkeit des steinkohlenbergbaues . fortführung der

Reichweite 3

zusammengefaßte ziele , maßnahmen und kosten . zur besseren übersicht werden die ziele durch , maßnahmen und kosten des nordrhein-westfalen-programms 1975 zunächst kurzfassung der einzelnen abschnitte einzelnen abschnitte teil . 3 . die arbeit und wirtschaft . kernenergie hochtemperaturreaktoren mit unmittelbar angeschlossener heliumgasturbine ; größerer anteil prozentuale anteil der kernenergie der stromerzeugung ; anwendung von prozeßwärme insbesondere zur kohlevergasung , erzverhüttung und herstellung chemischer rohstoffe . bau eines 300-mwe-hochtemperaturreaktors als prototyp und projektierung eines leistungsreaktors mit heliumgasturbine ; errichtung eines sonderforschungsbereiches " prozeßwärme " an der kernforschungsanlage jülich . bergbauanpassung steigerung der leistung und wettbewerbsfähigkeit des steinkohlenbergbaues . fortführung der

5.3.2. Levenshtein-Distanz

Zur Überprüfung der Dekodierergebnisse wird die Levenshtein-Distanz (Vgl. [Lev66]) verwendet. Die Levenshtein-Distanz entspricht in angewendeten Fall der Anzahl der Editier-Vorgänge, die man an einem Text vornehmen muss, um einen bestimmten anderen Text zu erhalten. Diese Editier-Vorgänge umfassen das Einfügen, Löschen und Ersetzen von Wörtern. Verglichen werden die Ursprungstexte dabei mit den jeweils durch Kodierung und Dekodierung entstandenen Ausgabetexten. Je niedriger die Distanz, umso mehr entspricht der Ausgabetext dem Ursprungstext.

Die Verwendung unterschiedlicher Reichweiten bei der Prädiktion führt zu deutlichen Unterschieden bei der Levenshtein-Distanz. Tabelle 22 zeigt die Ergebnisse für 10 der 35 benutzen Texte bei Benutzung von nur einer Prädiktionsmatrix der Reichweite 1, bei Benutzung der Reichweiten 1 und 2, sowie bei Benutzung aller 3 Reichweiten jeweils in 5 Ebenen. „Nr“ entspricht der Nummer des Textes im LIMAS-Korpus.

Die durchschnittliche Wortanzahl pro Text beträgt 2328 und die durchschnittliche Levenshtein-Distanz bei Benutzung von einer einzigen Verbindungsmatrix pro Ebene beträgt 1637 bei einem Tiefstwert von 1229 und einem Höchstwert von 2359. Dadurch wird deutlich wie unbrauchbar diese im Hinblick auf Textrekonstruktion ist. Eine drastische Verbesserung ist zu erkennen, wenn man zusätzlich Matrizen der Reichweite 2 verwendet. Die durchschnittliche Levenshtein-Distanz beträgt nur noch 457 bei einem Tiefstwert von 268 und einem Höchstwert von 771. Die Verwendung von allen Reichweitenmatrizen ergibt einen Durchschnittswert von 199 für die Levenshtein-Distanz. Der Tiefstwert beträgt 115 und der Höchstwert 379.

Tabelle 2: Levenshtein-Distanz für 10 Texte des LIMAS-Korpus bei Benutzung verschiedener maximaler Prädiktionsreichweiten für 5 Ebenen (Quelle: Eigene Darstellung)

Nr / Rw	21	39	52	63	70	72	79	186	209	301	Ø (~)
1	1242	1604	1317	2359	2008	1686	1519	1657	1229	1751	1637
2	268	374	341	771	680	490	467	375	355	445	457
3	146	140	121	379	376	196	149	145	115	227	199

Die Ergebnisse bei Benutzung von 2 Ebenen zeigt Tabelle 3. Der Durchschnittwert für die Reichweite 1 beträgt 997, für Reichweite 2 435 und für Reichweite 3 168. Die Werte für die Levenshtein-Distanz fallen also bei weniger Ebenen erwartungsgemäß niedriger aus, auf Kosten der Kompressionsrate.

Tabelle 3: Levenshtein-Distanz für 10 Texte des LIMAS-Korpus bei Benutzung verschiedener maximaler Prädiktionsreichweiten für 2 Ebenen (Quelle: Eigene Darstellung)

Nr / Rw	21	39	52	63	70	72	79	186	209	301	Ø (~)
1	744	1014	923	1285	1226	1008	969	973	853	979	997
2	210	350	323	753	644	474	461	371	325	435	435
3	60	104	109	373	284	178	129	129	99	213	168

Abbildung 8 zeigt die Ergebnisse in grafischer Form. Eine Verbesserung der Dekodierungsergebnisse durch Verwendung von mehr Prädiktionsmatrizen ist zu erkennen. Darüber hinaus ist eine weitere Verbesserung bei Verwendung von Matrizen der Reichweite 4, 5 usw. zu erwarten, wobei die Veränderungen tendenziell kleiner ausfallen werden.

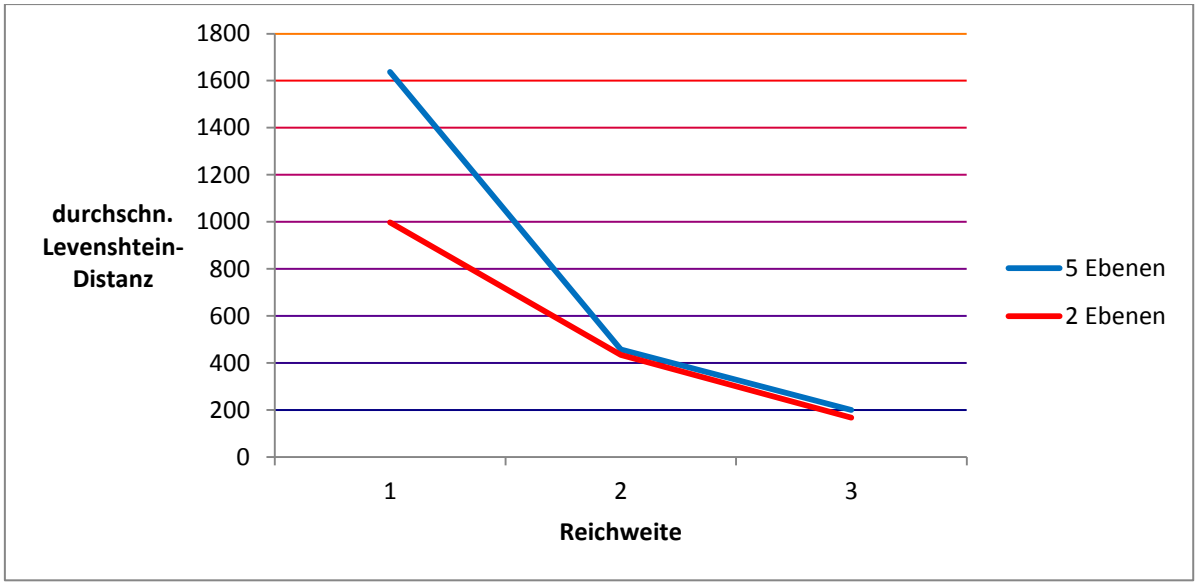

Abbildung 8: Durchschnittliche Levenshtein-Distanz bei verschiedener Reichweite und Ebenenanzahl (Quelle: Eigene Darstellung)

5.3.3. Vergleich Multi-Level/Single-Level

Da es sich um ein Kodierverfahren handelt, spielt abgesehen von der Levenshtein-Distanz auch die Kompressionsrate eine wichtige Rolle.

Aufgrund des Prinzips der hierarchischen Kodierung führt die Verwendung von mehr oder weniger Ebenen zu unterschiedlichen Ergebnissen bei der Kompressionsrate. Dabei verkürzt sich der kodierte Text pro Ebene jeweils ca. um die Hälfte. Dass es nicht genau die Hälfte ist liegt an den Startwörtern, die in jeder Ebene kodiert werden. Die Anzahl der Bitstellen der kodierten Texte bei Verwendung von 1 bzw. 5 Ebenen zeigt die Tabelle 4 für 8 verschiedene Texte.

Tabelle 4: Anzahl der Bit-Stellen verschiedener kodierter Texte bei Verwendung von 5 Ebenen im Vergleich zur nicht-hierarchischen Kodierung (Quelle: Eigene Darstellung)

Nr / Eb	21	39	52	63	70	72	79	186	Ø (~)
Wörter	2230	2306	2387	2299	2228	2354	2267	2411	2310
1	33454	34594	35809	34489	33424	35314	34009	36169	34658
5	2164	2239	2314	2224	2164	2284	2194	2329	2239
%	6,47	6,47	6,46	6,45	6,47	6,47	6,45	6,44	6,46

Die Anzahl der Bit-Stellen bei 5 Ebenen sinkt im Vergleich zur Kodierung auf einer Ebene auf ca. 6,46% bei ca. 2310 Wörtern. Je mehr Wörter der Text enthält, umso weniger spielen die Startwörter bei der Kompressionsrate eine Rolle. Sie würde sich bei 5 Ebenen dem Wert 6,25% nähern, also $\frac{1}{16}$.

Da beim Kodieren mit einer Ebene alle Wörter mit jeweils 15 bit kodiert werden, ohne Redundanz zu entfernen, ist bei der Dekodierung auch keine Prädiktion notwendig. Dementsprechend beträgt die Levenshtein-Distanz zwischen dem originalen und dem dekodierten Text immer 0.

6. Mögliche Verbesserungen/Erweiterungen

Stammformen der Wörter

Im verwendeten Programm werden z.B. die Wörter „gehen“, „ging“, „gegangen“ und „geht“ wie völlig unterschiedliche Wörter behandelt. Jedes Wort hat seine eigenen Vorgänger und Nachfolger. Es wäre aber auch möglich, diese Wörter unter der Stammform „gehen“ zusammenzufassen, wobei es dann wiederum ein Extrasystem geben müsste, um beim Dekodieren wieder von der Stammform zum richtigen Wort zu gelangen. (Vgl. [Rie02])

Segmentierung, zusammenfassen zweier Wörter

Bei zwei Wörtern, die sehr häufig aufeinanderfolgen, wie z.B. „die Katze“, kann es sinnvoll sein, diese als ein Wort zu behandeln. In den Matrizen sind also Vorgänger und Nachfolger von „die Katze“ eingetragen. Dadurch erhöht sich zwar die Anzahl der Wörter, aber die Dekodierergebnisse werden vermutlich verbessert. Das gleiche Prinzip ist auch mit mehr als zwei Wörtern denkbar. (Vgl. [Rie02])

Mitüberlieferung von benötigten zusätzlichen Verbindungen in den vorhandenen Matrizen

Die Verbindungsmatrizen im verwendeten Programm sind starr. Sie verändern sich nicht beim kodieren neuer Texte. Allerdings ist es bei neuen Texten wahrscheinlich, dass vor allem die ranghohen Wörter Verbindungen haben, welche in den Lerntexten und damit in der Verbindungsmatrizen so nicht vorkommen. Dadurch wird es beim dekodieren schwierig, den korrekten Text zu prädizieren. Aus diesem Grund wurden in dieser Arbeit auch nur Texte kodiert, welche bereits Teil des Lernkorpus waren.

Umgehen kann man das, indem man beim Kodieren die relevanten Verbindungen in die Matrizen einträgt und diese Veränderungen auch als Teil des Kodes an die Dekodiermaschine weiterleitet. (Vgl. [Hil05], S. 62-63)

Trennen nach Sätzen:

Betrachtet man Punkte als Wörter und damit die unterschiedlichen Sätze nicht separat, wird es häufig passieren, dass Satzende eines Satzes und Satzanfang des nächsten Satzes zu einem Metawort in der nächsthöheren Ebene verschmelzen. Das neue Metawort bzw. die Vorgänger-Nachfolger-Verbindung ist in diesem Fall nicht zwangsläufig sinnvoll. Um das zu umgehen können sogenannte Regieanweisungen verwendet werden. (Vgl. [Hil05], S. 39)

7. Bewertung und Ausblick

Die Sprachmaschine soll in Zukunft das menschliche Denken simulieren können, davon ist man aber noch weit entfernt. Man konnte die bereits gewonnenen Erkenntnisse und Ansätze aber nutzen um ein Kodierverfahren zu entwickeln, welches hervorragende Ergebnisse bei der Kompressionsrate liefert.

Ziel dieser Arbeit war es, die grundlegenden Algorithmen dieses Kodierverfahrens in einem MATLAB-Programm zu realisieren und auf ihre Funktionsfähigkeit und Richtigkeit zu überprüfen, sowie das Zipfsche Gesetz zu bestätigen. Die von Programm gelieferten Ergebnisse entsprachen den Erwartungen. Der Dekodieralgorithmus liefert lesbare Texte, prädiziert aus wenigen Worten der 5. Ebene.

Mit dem im Rahmen dieser Arbeit erstellten Programm ist noch keine sinnvolle alltägliche Anwendung möglich. Es ist aber ein Fundament, auf dem aufgebaut werden kann. Dazu wurden auch bereits Verbesserungsvorschläge vorgestellt.

Literaturverzeichnis

[Bur98] **Burschel, Horst-Dieter (1998):** Die meßtechnische Ermittlung von Assoziationen zwischen Worten in kohärentem Text und ihre Nutzung bei Prädiktionen. Dissertation, TU Darmstadt

[Dud09] **Duden (2009):** Zum Umfang des deutschen Wortschatzes, Internetadresse: http://www.duden.de/sprachwissen/sprachratgeber/zum-umfang-des-deutschen-wortschatzes

[Göb07] **Göbel, Jürgen (2007):** Informationstheorie und Codierungsverfahren. Berlin: VDE Verlag

[Hil05] **Hilberg, Wolfgang (2005):** Denken wie ein Mensch, Netzwerk-Strukturen eines neuronalen Sprachsystems mit einer Architektur nach menschlichen Vorbild. Groß-Bieberau/Odenwald: Verlag für Sprache und Technik

[Hil08] **Hilberg, Wolfgang (2008):** Sprache und Denken in neuronalen Netzen. Groß-Bieberau/Odenwald: Verlag für Sprache und Technik

[Hil12] **Hilberg, Wolfgang (2012):** Wie denkt das Gehirn?, Die Lösung des alten Rätsels durch neuronale Repräsentation von Sprache. Groß-Bieberau/Odenwald: Verlag für Sprache und Technik

[Lev66] **Levenshtein, Vladimir (1966):** Binary Codes Capable Of Correcting Deletions, Insertions and Reversals. In: Soviet Physics Doklady, 10(8), S. 707–710

[Lyr02] **Lyre, Holger (2002):** Informationstheorie. München: Wilhelm Fink Verlag

[Rie02] **Ries, Thomas (2002):** Über Möglichkeiten einer maschinellen Nacherzählung mit einem konnektionistischen System aus neuronalen Sprachnetzwerk. Dissertation, TU Darmstadt

[Sch12] **Schönfeld, Dagmar; Klimant, Herbert und Piotraschke, Rudi (2012):** Informations- und Kodierungstheorie. 4. Auflage, Wiesbaden: Springer Vieweg

Erklärung

Der Verfasser erklärt, dass er die vorliegende Arbeit selbständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt hat. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Cottbus, 27.05.15 _____