Technische Universität Berlin

Quality and Usability Lab

Part-of-Speech Tagging with Neural Networks for a Conversational Agent

Master Thesis

Master of Science (M.Sc.)

Author	Andreas Müller
Major	Computer Engineering
Matriculation No.	333471

Date	???
1st supervisor	Prof. Dr.-Ing. Sebastian Möller
2nd supervisor	Prof. Dr. ???

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht.

Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

Berlin, den 16. November 2017

_____

Unterschrift

Zusammenfassung

Zur Verarbeitung großer Textmengen und automatisierter Modellierung von Sprache gibt es viele Ansätze. Dabei erreicht *Deep Learning* mithilfe neuronaler Netzwerke basierend auf englischer Sprache heutzutage nicht nur hervorragende Ergebnisse, sondern kann Modelle mithilfe geeigneter Architekturen auch unüberwacht trainieren.

Diese Arbeit beschreibt die Modellierung natürlicher Sprache mithilfe von neuronalen Netzwerken basierend auf deutschen Textkorpora. Es soll eine Aussage darüber getroffen werden, wie gut die resultierenden Wort-Vektoren die deutsche Sprache repräsentieren. Dazu werden unter Parametervariation verschiedene Modelle trainiert und mithilfe in dieser Arbeit entwickelter Test-Sets evaluiert. Das resultierende Modell mit den besten Evaluationsergebnissen dient anschließend als Grundlage für die Darstellung komplexer Wortzusammenhänge und zur Visualisierung sprachlicher Konzepte.

Abstract

There are several approaches to process large amounts of text and to model language automatically. *Deep Learning* not only achieves outstanding results in modeling the English language with the help of neural networks, but is also able to train models unsupervised with appropriate architectures.

This thesis describes the modeling of natural language using neural networks with German text corpora. The precision of the resulting German word embeddings is analyzed. Various models are trained under parameter variation and evaluated with the help of test sets, that are generated within this work. The resulting model with optimal parameter configuration is then used for complex semantic word connections and visualization of linguistic concepts.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

NLP	*Natural Language Processing* (linguistische Datenverarbeitung)
IR	*Information Retrieval* (Informationsrückgewinnung)
SGD	*Stochastic Gradient Descent*
NNLM	*(Feedforward) Neural Net Language Model*
RNNLM	*Recurrent Neural Net Language Model*
CBOW	*Continous Bag-of-Words*
PCA	*Principal Component Analysis* (Hauptkomponentenanalyse)

1 Einleitung

Die Analyse und Verarbeitung von Texten natürlicher Sprache stellt die Grundlage für zahlreiche Anwendungen der linguistischen Datenverarbeitung (engl. *Natural Language Processing*, NLP) dar und ist heute ein zentrales Forschungsgebiet von Internetfirmen wie Facebook[1], Google[2] oder Yahoo[3], denen große Mengen an Textdaten zur Verfügung stehen.

NLP möchte im Allgemeinen die natürliche Sprache algorithmisch verarbeiten, sodass automatisiert unstrukturierte Daten in geordnete Informationen umgewandelt und gewünschte Informationen extrahiert werden können. Dieser Vorgang wird als *Information Retrieval* (IR) bezeichnet. Teil des IR ist die Information Extraction, welche gezielt Informationen einer bestimmten Vorgabe gewinnen möchte. Um natürliche Sprache automatisiert zu verarbeiten, müssen Wörter in ein geeignetes maschinenlesbares Format umgewandelt werden. Wort-Vektoren stellen eine Möglichkeit dar, Wörter numerisch zu beschreiben und ihren sprachlichen Zusammenhang durch die Beziehung der Vektoren in einem beschränkten Vektorraum abzubilden (Bengio et al., 2003 [4]).

Sprach-Modellierung kann mithilfe verschiedener Ansätze realisiert werden. Zum einen gibt es einfache Modelle wie *Bag-of-Words* (BOW, vgl. **??**) oder N-Gramme (vgl. **??**). Die Größe des Vokabulars ist hier maßgebend für die Dimension der resultierenden Wort-Vektoren, sodass webbasierte Korpora (wie beispielsweise die freie Enzyklopädie Wikipedia) aufgrund ihrer Größe auch bei hoher Rechenleistung sehr lange Berechnungszeiten verursachen. Da z.B. beim BOW-Modell jedes neue Wort im Vokabular jedem Wort-Vektor eine weitere Dimension hinzufügt, verlieren folglich die Wort-Vektoren bezüglich des repräsentierten Wortes zunehmend an Bedeutung[4].

1 Facebook veröffentlichte 2013 einen wissenschaftlichen Artikel zu *Unicorn* (Curtiss et al. [7]), eines Indexing-System für Facebook's Graph Search, welche es ermöglicht, Suchanfragen der Art *„Restaurants in San Francisco liked by people from Beijing"* zu stellen.

2 Google kündigte Ende September 2013 *Hummingbird* an (Shapiro et al. [24]), einen Algorithmus zur effizienteren Verarbeitung und Sortierung des Suchindex. Damit soll die Bedeutung eines Satzes besser verstanden werden, um präzisere Suchergebnisse komplexer Suchanfragen zu erreichen.

3 Yahoo kaufte Ende 2013 *SkyPhrase* (Miners et al. [16]), ein NLP Start-Up, um die Verarbeitung und Resultate von Benutzeranfragen in vielen Yahoo-Anwendungen zu verbessern.

4 Dieses Phänomen ist als *Curse of Dimensionality* – Fluch der Dimensionalität bekannt. Der Begriff wurde 1961 erstmals verwendet von R.E. Bellman [2].

Zum anderen gibt es beim *Deep Learning* den Ansatz der neuronalen Netzwerke, welche die Wort-Vektoren durch mehrschichtige Abstraktionen ohne Vorgabe trainieren können. Es wurde gezeigt, dass diese mithilfe von neuronalen Netzwerken gelernten Wort-Vektoren sprachliche Merkmale sehr gut abbilden konnten (Bengio et al., 2003 [4]). Dadurch war es möglich, komplexere Aufgaben zu lösen wie beispielsweise *Relation Detection*, *Relation Classification* (Zeng et al., 2014 [27]) oder *Sentiment Analysis* (Socher et al., 2013 [25]).

Viele Modelle wurden bisher fast ausschließlich für die englische Sprache trainiert, sodass es für die deutsche Sprache kaum Modelle gibt, mit deren Hilfe man Aussagen darüber treffen könnte, wie präzise deutsche Wort-Vektoren sind und welche Parameter genauere Ergebnisse auf einem deutschen Korpus erzielen. Aufgrund der sprachlichen Unterschiede zwischen Deutsch und Englisch können Ergebnisse englischer Sprachmodelle nicht ohne Weiteres für die deutsche Sprache übernommen werden. So gibt es im Deutschen beispielsweise männliche, weibliche und sächliche Artikel, wobei sich biologisches und grammatikalisches Geschlecht unterscheiden können. Gegenstände sind nicht wie im Englischen ausschließlich sächlich, sondern haben oft ein Geschlecht, wie „**der** Stuhl“ oder „**die** Lampe“ (***the*** *chair*, ***the*** *lamp*). Außerdem gibt es Unterschiede in der Wort-Reihenfolge: Während englische Sätze und Nebensätze immer dem Schema ‘Subjekt - Prädikat - Objekt’ folgen, so steht in deutschen Nebensätzen das Verb an letzter Stelle, z.B. „Ich weiß, dass **sie** das Buch **liest**.“ (*I know that* ***she reads*** *the book.*), wobei dem Verb auch längere Beschreibungen vorangestellt werden können. Werden Sprachmodelle aufgrund einer bestimmten Anzahl nebeneinander stehender Wörter (Fenster) trainiert, so ergeben sich für beide Sprachen schon wegen der Unterschiede in der Wort-Reihenfolge verschiedene Ergebnisse.

Ziel dieser Arbeit ist es, die durch Variation verschiedener Parameter beim Modelltraining entstandenen deutschen Wort-Vektoren zu analysieren und zu evaluieren. Dazu wird ein Toolkit für Korpuserstellung, Training, Evaluation und Visualisierung entwickelt und für die Evaluation deutsche Test-Sets generiert. Dieses Toolkit, das finale Sprachmodell und die Test-Sets stehen anschließend als Grundlage für weiterführende Anwendungen deutscher Wort-Vektoren zur Verfügung.

1 Einleitung

Anknüpfen soll diese Arbeit an die Ergebnisse des DIMA Projektes „Exploring semantic word similarities in German News Articles“ [1], welches sich mit unüberwachtem Clustering deutscher Nachrichtenartikel beschäftigte. Basierend auf einem Korpus von 3 Millionen Nachrichtenartikeln wurden in diesem Projekt bereits verschiedene Modelle unter Parameter-Variation mithilfe eines neuronalen Netzwerks auf deutscher Sprache trainiert. Es wurde gezeigt, dass das Trainieren von Wort-Vektoren auf einem Korpus deutscher Sprache prinzipiell funktioniert, sodass darauf aufbauend in dieser Arbeit das beschriebene Toolkit zur Bewertung deutscher Modelle implementiert werden kann.

Diese Arbeit ist in 6 Kapitel gegliedert. In **??** werden zunächst die Grundlagen zu Wort-Vektoren allgemein und maschinellem Lernen in der linguistischen Datenverarbeitung dargestellt. Anschließend wird in **??** genauer auf den Aufbau neuronaler Netzwerke und darauf basierenden Architekturen und Modellen eingegangen. Das Training der Modelle und die daraus resultierenden verbesserten Wort-Vektoren werden hier erläutert. Die Umsetzung des praktischen Teils, der Implementierung eines Toolkits zur Korpora-Erstellung, Modelltraining und Evaluation trainierter Modelle, ist in **??** dargestellt. Dabei werden durch Parametervariation verschiedene Modelle spezifiziert und die Erstellung von Test-Sets erläutert. In **??** werden die Ergebnisse der trainierten Modelle sowie das daraus resultierende optimale Modell ausgewertet. Abschließend werden die wichtigsten Erkenntnisse in Kapitel 2 zusammengefasst.

Da es in dieser Arbeit um die Modellierung deutscher Sprache geht, werden weitestgehend die üblicherweise verwendeten deutschen Entsprechungen englischer Fachbegriffe verwendet.

2 Fazit

Zur Verarbeitung großer Textmengen und automatisierter Modellierung von Sprache gibt es viele Ansätze. Dabei erreicht *Deep Learning* mithilfe neuronaler Netzwerke basierend auf englischer Sprache heute nicht nur hervorragende Ergebnisse, sondern kann Modelle mithilfe von Architekturen wie Skip-Gram oder CBOW auch unüberwacht trainieren.

Während das Trainieren von Wort-Vektoren bisher hauptsächlich für die englische Sprache durchgeführt wurde, hat diese Arbeit gezeigt, dass eine solche Sprachmodellierung auch für die deutsche Sprache funktioniert. Es wurden mithilfe eines dafür entwickelten Toolkits unter Parametervariation insgesamt 25 unterschiedliche Modelle und daraus resultierend ein optimales Modell trainiert[1], welches bei der Evaluierung syntaktischer Merkmale ein ähnliches Niveau erreicht, wie ein vergleichbares System (vgl. **??**). Dieses Training war dabei aufgrund einer recheneffizienten Implementierung von Skip-Gram und CBOW auch auf großen Korpora mit einem normalen Heimrechner (vgl. **??**) möglich. Beliebige und große Korpora konnten verwendet werden, da das Training unüberwacht stattfand und deshalb keine gelabelten Eingabedaten nötig waren, sondern ausschließlich Text, wie er im normalen Sprachgebrauch auftritt.

Mit der aus der Analyse der Evaluationsergebnisse gefundenen optimalen Parameterkofiguration konnten die im Rahmen dieser Arbeit besten deutschen Wort-Vektoren trainiert werden. Mit diesen Wort-Vektoren war es möglich, syntaktische und semantische Wort-Zusammenhänge darzustellen, komplexe Wortbeziehungen in ihrer Bedeutung korrekt zuordnen (vgl. **??**) und Merkmale mithilfe von PCA zu visualisieren.

Die in dieser Arbeit trainierten Wort-Vektoren können nun als Grundlage für weiterführende Forschungen und Anwendungen der Sprachmodellierung für die deutsche Sprache dienen, wie z.B. *Relation Detection* bzw. *Relation Classification* (das Modellieren und anschließende Klassifizieren der Beziehung von Wörtern zueinander) oder *Sentiment Analysis* (das Erkennen positiver oder negativer Aussagen).

1 Dieses deutsche Sprachmodell, das entwickelte Toolkit und die generierten Test-Set Fragen stehen für weiterführende Anwendungen zum freien Download zur Verfügung. Müller, Andreas (o.J.), URL: http://devmount.github.io/GermanWordEmbeddings (Stand: 30.06.2015)

A Anhang

A.1 Training

Listing A.1: Liste der im Training verwendeten Deutschen Stoppwörter des Natural Language Toolkit (NLTK)

```
aber
alle
allem
allen
aller
alles
als
also
am
an
ander
andere
anderem
anderen
anderer
anderes
anderm
andern
anderr
anders
auch
auf
aus
bei
bin
bis
bist
da
damit
dann
der
den
des
dem
die
das
daß
derselbe
derselben
denselben
desselben
demselben
dieselbe
dieselben
dasselbe
dazu
dein
deine
deinem
deinen
deiner
deines
denn
derer
dessen
dich
dir
du
dies
diese
diesem
diesen
dieser
dieses
doch
dort
durch
ein
eine
einem
einen
einer
eines
einig
einige
einigem
einigen
einiger
einiges
einmal
er
ihn
ihm
es
etwas
euer
eure
eurem
euren
eurer
eures
für
gegen
gewesen
hab
habe
haben
hat
hatte
hatten
hier
hin
hinter
ich
mich
mir
ihr
ihre
ihrem
ihren
ihrer
ihres
euch
im
in
indem
ins
ist
jede
jedem
jeden
jeder
jedes
jene
jenem
jenen
jener
jenes
jetzt
kann
kein
keine
keinem
keinen
keiner
keines
können
könnte
machen
man
manche
manchem
manchen
mancher
manches
mein
meine
meinem
meinen
meiner
meines
mit
muss
musste
nach
nicht
nichts
noch
nun
nur
ob
oder
ohne
sehr
sein
seine
seinem
seinen
seiner
seines
selbst
sich
sie
ihnen
sind
so
solche
solchem
solchen
solcher
solches
soll
sollte
sondern
sonst
über
um
und
uns
unse
unsem
unsen
unser
unses
unter
viel
vom
von
vor
während
war
waren
warst
was
weg
weil
weiter
welche
welchem
welchen
welcher
welches
wenn
werde
werden
wie
wieder
will
wir
wird
wirst
wo
wollen
wollte
würde
würden
zu
zum
zur
zwar
zwischen
```

A.2 Test-Sets

Listing A.2: Syntaktische Analogiefragen (Auszug): Eine Frage (bestehend aus zwei Wortpaaren) pro Zeile, Zeile beginnend mit Doppelpunkt ist das Label des Test-Musters.

```
: nouns: SI/PL
Abbildung Abbildungen Ergebnis Ergebnisse
Abbildung Abbildungen Name Namen
Abbildung Abbildungen Person Personen
Abbildung Abbildungen Ziel Ziele
Abbildung Abbildungen Nacht Nächte
Absatz Absätze Wohnung Wohnungen
Absatz Absätze Boden Böden
Absatz Absätze Straße Straßen
Absatz Absätze Erfahrung Erfahrungen
Absatz Absätze Schule Schulen
[...]
: nouns: PL/SI
Abbildungen Abbildung Schulen Schule
Abbildungen Abbildung Orte Ort
Abbildungen Abbildung Minuten Minute
Abbildungen Abbildung Erfahrungen Erfahrung
Abbildungen Abbildung Bereiche Bereich
Absätze Absatz Herren Herr
Absätze Absatz Nächte Nacht
Absätze Absatz Universitäten Universität
Absätze Absatz Zeiten Zeit
```

A Anhang

```
Absätze Absatz Familien Familie
[...]
: adjectives: GR/KOM
ähnlich ähnlicher faul fauler
ähnlich ähnlicher ruhig ruhiger
ähnlich ähnlicher nett netter
ähnlich ähnlicher eckig eckiger
ähnlich ähnlicher alt älter
alt älter bestimmt bestimmter
alt älter früh früher
alt älter ähnlich ähnlicher
alt älter lang länger
alt älter wichtig wichtiger
[...]
: adjectives: KOM/GR
ähnlicher ähnlich eckiger eckig
ähnlicher ähnlich schärfer scharf
ähnlicher ähnlich trauriger traurig
ähnlicher ähnlich wahrscheinlicher wahrscheinlich
ähnlicher ähnlich langsamer langsam
älter alt genauer genau
älter alt voller voll
älter alt krümmer krumm
älter alt sauberer sauber
älter alt unterschiedlicher unterschiedlich
[...]
: adjectives: GR/SUP
ähnlich ähnlichste kalt kälteste
ähnlich ähnlichste schlank schlankeste
ähnlich ähnlichste nett netteste
ähnlich ähnlichste spannend spannendste
ähnlich ähnlichste schön schönste
alt älteste deutlich deutlichste
alt älteste hart härteste
alt älteste sauber sauberste
alt älteste schmal schmalste
alt älteste nett netteste
[...]
: adjectives: SUP/GR
ähnlichste ähnlich persönlichste persönlich
ähnlichste ähnlich langsamste langsam
ähnlichste ähnlich frischste frisch
ähnlichste ähnlich vollste voll
ähnlichste ähnlich wertvollste wertvoll
älteste alt liebste lieb
älteste alt häufigste häufig
älteste alt weiteste weit
älteste alt kürzeste kurz
älteste alt sonnigste sonnig
[...]
: adjectives: KOM/SUP
ähnlicher ähnlichste größer größte
ähnlicher ähnlichste sonniger sonnigste
ähnlicher ähnlichste schrecklicher schrecklichste
ähnlicher ähnlichste schmutziger schmutzigste
ähnlicher ähnlichste schneller schnellste
älter älteste schärfer schärfste
älter älteste gerader geradeste
älter älteste später späteste
älter älteste gerader geradeste
älter älteste fleißiger fleißigste
[...]
: adjectives: SUP/KOM
ähnlichste ähnlicher schnellste schneller
ähnlichste ähnlicher bestimmteste bestimmter
ähnlichste ähnlicher leichteste leichter
ähnlichste ähnlicher richtigste richtiger
ähnlichste ähnlicher schwierigste schwieriger
älteste älter müdeste müder
älteste älter stärkste stärker
älteste älter stillste stiller
älteste älter frischste frischer
älteste älter ruhigste ruhiger
[...]
: verbs (pres): INF/1SP
ändern ändere schauen schaue
ändern ändere wissen weiß
ändern ändere suchen suche
ändern ändere kommen komme
ändern ändere denken denke
arbeiten arbeite erhalten erhalte
arbeiten arbeite tragen trage
arbeiten arbeite erwarten erwarte
arbeiten arbeite fühlen fühle
arbeiten arbeite wohnen wohne
[...]
: verbs (pres): 1SP/INF
ändere ändern brauche brauchen
ändere ändern werde werden
ändere ändern schaffe schaffen
ändere ändern zeige zeigen
ändere ändern bleibe bleiben
arbeite arbeiten sehe sehen
arbeite arbeiten liege liegen
arbeite arbeiten stelle stellen
arbeite arbeiten muss müssen
arbeite arbeiten schließe schließen
[...]
: verbs (pres): INF/2PP
ändern ändert verstehen versteht
ändern ändert fehlen fehlt
ändern ändert erwarten erwartet
ändern ändert entstehen entsteht
ändern ändert wachsen wachst
arbeiten arbeitet fragen fragt
arbeiten arbeitet sprechen sprecht
arbeiten arbeitet sitzen sitzt
arbeiten arbeitet suchen sucht
arbeiten arbeitet fallen fallt
[...]
: verbs (pres): 2PP/INF
ändert ändern erkennt erkennen
ändert ändern geltet gelten
ändert ändern geltet gelten
ändert ändern vergleicht vergleichen
ändert ändern lebt leben
arbeitet arbeiten erscheint erscheinen
arbeitet arbeiten versucht versuchen
arbeitet arbeiten bekommt bekommen
arbeitet arbeiten ergebt ergeben
arbeitet arbeiten sitzt sitzen
[...]
: verbs (pres): 1SP/2PP
ändere ändert soll sollt
ändere ändert vergleiche vergleicht
ändere ändert treffe trefft
ändere ändert wohne wohnt
ändere ändert rede redet
arbeite arbeitet laufe lauft
arbeite arbeitet entwickle entwickelt
arbeite arbeitet gehöre gehört
arbeite arbeitet rede redet
arbeite arbeitet versuche versucht
[...]
: verbs (pres): 2PP/1SP
ändert ändere besteht bestehe
ändert ändere kommt komme
ändert ändere bietet biete
ändert ändere gewinnt gewinne
ändert ändere arbeitet arbeite
arbeitet arbeite fehlt fehle
arbeitet arbeite vergleicht vergleiche
```

```
arbeitet arbeite erkennt erkenne
arbeitet arbeite führt führe
arbeitet arbeite seht sehe
[...]
: verbs (past): INF/3SV
ändern änderte mögen mochte
ändern änderte gehören gehörte
ändern änderte entwickeln entwickelte
ändern änderte gelten galt
ändern änderte machen machte
arbeiten arbeitete ändern änderte
arbeiten arbeitete bleiben blieb
arbeiten arbeitete wollen wollte
arbeiten arbeitete kommen kam
arbeiten arbeitete brauchen brauchte
[...]
: verbs (past): 3SV/INF
änderte ändern musste müssen
änderte ändern spielte spielen
änderte ändern bot bieten
änderte ändern studierte studieren
änderte ändern redete reden
arbeitete arbeiten betraf betreffen
arbeitete arbeiten hieß heißen
arbeitete arbeiten saß sitzen
arbeitete arbeiten legte legen
arbeitete arbeiten fiel fallen
[...]
: verbs (past): INF/3PV
ändern änderten leben lebten
ändern änderten halten hielten
ändern änderten sagen sagten
ändern änderten erkennen erkannten
ändern änderten gelten galten
arbeiten arbeiteten sagen sagten
arbeiten arbeiteten können konnten
arbeiten arbeiteten lernen lernten
arbeiten arbeiteten müssen mussten
arbeiten arbeiteten gehören gehörten
[...]
: verbs (past): 3PV/INF
änderten ändern verbanden verbinden
änderten ändern handelten handeln
änderten ändern trafen treffen
änderten ändern lasen lesen
änderten ändern erschienen erscheinen
arbeiteten arbeiten folgten folgen
arbeiteten arbeiten saßen sitzen
arbeiteten arbeiten wussten wissen
arbeiteten arbeiten schrieben schreiben
arbeiteten arbeiten halfen helfen
[...]
: verbs (past): 3SV/3PV
änderte änderten sprach sprachen
änderte änderten verband verbanden
änderte änderten fand fanden
änderte änderten zeigte zeigten
änderte änderten nahm nahmen
arbeitete arbeiteten fiel fielen
arbeitete arbeiteten dachte dachten
arbeitete arbeiteten schrieb schrieben
arbeitete arbeiteten trug trugen
arbeitete arbeiteten dachte dachten
[...]
: verbs (past): 3PV/3SV
änderten änderte interessierten interessierte
änderten änderte lebten lebte
änderten änderte bildeten bildete
änderten änderte schlossen schloss
änderten änderte brauchten brauchte
arbeiteten arbeitete mochten mochte
arbeiteten arbeitete hatten hatte
arbeiteten arbeitete setzten setzte
arbeiteten arbeitete handelten handelte
arbeiteten arbeitete fragten fragte
```

Listing A.3: Thematische Analogiefragen (Auszug): Eine Frage (bestehend aus zwei Wortpaaren) pro Zeile.

```
China Yuan Deutschland Euro
China Yuan Dänemark Krone
China Yuan England Pfund
China Yuan Japan Yen
China Yuan Russland Rubel
China Yuan USA Dollar
Deutschland Euro Dänemark Krone
Deutschland Euro England Pfund
Deutschland Euro Japan Yen
Deutschland Euro Russland Rubel
Deutschland Euro USA Dollar
[...]
Berlin Deutschland Bern Schweiz
Berlin Deutschland Hanoi Vietnam
Berlin Deutschland Helsinki Finnland
Berlin Deutschland Kairo Ägypten
Berlin Deutschland Kiew Ukraine
Berlin Deutschland London England
Berlin Deutschland Madrid Spain
Berlin Deutschland Melbourne Australien
Berlin Deutschland Moskau Russland
Berlin Deutschland Oslo Norwegen
Berlin Deutschland Ottawa Kanada
Berlin Deutschland Paris Frankreich
Berlin Deutschland Rom Italien
Berlin Deutschland Stockholm Schweden
Berlin Deutschland Teheran Iran
Berlin Deutschland Tokio Japan
Berlin Deutschland Washington USA
[...]
England Europa Frankreich Europa
England Europa Griechenland Europa
England Europa Indien Asien
England Europa Italien Europa
England Europa Kanada Nordamerika
England Europa Polen Europa
England Europa USA Nordamerika
England Europa Vietnam Asien
England Europa Ägypten Afrika
Frankreich Europa Griechenland Europa
Frankreich Europa Indien Asien
Frankreich Europa Italien Europa
Frankreich Europa Kanada Nordamerika
Frankreich Europa Polen Europa
Frankreich Europa USA Nordamerika
Frankreich Europa Vietnam Asien
Frankreich Europa Ägypten Afrika
[...]
Frankreich Französisch Griechenland Griechisch
Frankreich Französisch Italien Italienisch
Frankreich Französisch Japan Japanisch
Frankreich Französisch Korea Koreanisch
```

```
Frankreich Französisch Norwegen Norwegisch
Frankreich Französisch Polen Polnisch
Frankreich Französisch Russland Russisch
Frankreich Französisch Schweden Schwedisch
Frankreich Französisch Spanien Spanisch
Frankreich Französisch Ukraine Ukrainisch
Griechenland Griechisch Italien Italienisch
Griechenland Griechisch Japan Japanisch
Griechenland Griechisch Korea Koreanisch
Griechenland Griechisch Norwegen Norwegisch
Griechenland Griechisch Polen Polnisch
Griechenland Griechisch Russland Russisch
Griechenland Griechisch Schweden Schwedisch
Griechenland Griechisch Spanien Spanisch
Griechenland Griechisch Ukraine Ukrainisch
[...]
Elisabeth Königin Charles Prinz
Android Google iOS Apple
Android Google Windows Microsoft
iOS Apple Windows Microsoft
[...]
Junge Mädchen König Königin
Junge Mädchen Mann Frau
Junge Mädchen männlich weiblich
Junge Mädchen Neffe Nichte
Junge Mädchen Onkel Tante
Junge Mädchen Papa Mama
Junge Mädchen Partner Partnerin
Junge Mädchen Prinz Prinzessin
Junge Mädchen Vater Mutter
König Königin Mann Frau
König Königin männlich weiblich
König Königin Neffe Nichte
König Königin Onkel Tante
König Königin Papa Mama
König Königin Partner Partnerin
König Königin Prinz Prinzessin
König Königin Vater Mutter
Mann Frau männlich weiblich
Mann Frau Neffe Nichte
Mann Frau Onkel Tante
Mann Frau Papa Mama
Mann Frau Partner Partnerin
Mann Frau Prinz Prinzessin
Mann Frau Vater Mutter
[...]
```

Listing A.4: Fragen zum inhaltlich nicht passenden Wort einer Wortreihe (Auszug): Eine Frage (bestehend aus drei zueinander passenden Wörtern und einem nicht passenden vierten Wort) pro Zeile.

```
August April September Jahr
August April September Monat
August April September Tag
August April September Stunde
August April September Minute
August April September Zeit
August April September Kalender
August April September Woche
August April September Quartal
August April September Uhr
Auto Motorrad Fahrrad Ampel
Auto Motorrad Fahrrad Fahrbahn
Auto Motorrad Fahrrad Fahrer
Auto Motorrad Fahrrad Fußgänger
Auto Motorrad Fahrrad Karte
Auto Motorrad Fahrrad Navigation
Auto Motorrad Fahrrad Polizei
Auto Motorrad Fahrrad Schild
Auto Motorrad Fahrrad Straße
Auto Motorrad Fahrrad Verkehr
Berlin München Frankfurt Amsterdam
Berlin München Frankfurt Brüssel
Berlin München Frankfurt Deutschland
Berlin München Frankfurt Indien
Berlin München Frankfurt Kopenhagen
Berlin München Frankfurt London
Berlin München Frankfurt Luxemburg
Berlin München Frankfurt Paris
Berlin München Frankfurt Washington
Berlin München Frankfurt Wien
Euro Rubel Yen Australien
Euro Rubel Yen China
Euro Rubel Yen Deutschland
Euro Rubel Yen England
Euro Rubel Yen Frankreich
Euro Rubel Yen Indien
Euro Rubel Yen Japan
Euro Rubel Yen Kanada
Euro Rubel Yen Russland
Euro Rubel Yen USA
Frankreich Deutschland England Afrika
Frankreich Deutschland England Amerika
Frankreich Deutschland England Asien
Frankreich Deutschland England Australien
Frankreich Deutschland England Brasilien
Frankreich Deutschland England China
Frankreich Deutschland England Europa
Frankreich Deutschland England Kanada
Frankreich Deutschland England Mexiko
Frankreich Deutschland England USA
Hase Hund Katze Baum
Hase Hund Katze Besitzer
Hase Hund Katze Elefant
Hase Hund Katze Essen
Hase Hund Katze Haus
Hase Hund Katze Mensch
Hase Hund Katze Tier
Hase Hund Katze Tierheim
Hase Hund Katze Wiese
Hase Hund Katze Zoo
Herz Lunge Leber Arzt
Herz Lunge Leber Blut
Herz Lunge Leber Fuß
Herz Lunge Leber Gesundheit
Herz Lunge Leber Hand
Herz Lunge Leber Kopf
Herz Lunge Leber Krankenhaus
Herz Lunge Leber Krankheit
Herz Lunge Leber Körper
Herz Lunge Leber Organ
Montag Mittwoch Freitag Jahr
Montag Mittwoch Freitag Monat
Montag Mittwoch Freitag Tag
Montag Mittwoch Freitag Stunde
Montag Mittwoch Freitag Minute
Montag Mittwoch Freitag Zeit
```

```
Montag Mittwoch Freitag Kalender
Montag Mittwoch Freitag Woche
Montag Mittwoch Freitag Wochentag
Montag Mittwoch Freitag Uhr
[...]
Twitter Facebook Instagram Android
Twitter Facebook Instagram App
Twitter Facebook Instagram Computer
Twitter Facebook Instagram Domain
Twitter Facebook Instagram Internet
Twitter Facebook Instagram iOS
Twitter Facebook Instagram Laptop
Twitter Facebook Instagram Microsoft
Twitter Facebook Instagram Netzwerk
Twitter Facebook Instagram Software
Windows Linux Android App
Windows Linux Android Computer
Windows Linux Android Gerät
Windows Linux Android Laptop
Windows Linux Android Programm
Windows Linux Android Rechner
Windows Linux Android Smartphone
Windows Linux Android Software
Windows Linux Android Tablet
Windows Linux Android Technik
```

Listing A.5: Analogiefragen zum Gegenteil eines Wortes (Auszug): Eine Frage (bestehend aus zwei Wortpaaren) pro Zeile.

```
Frage Antwort stark schwach
Frage Antwort viel wenig
Frage Antwort positiv negativ
Frage Antwort davor danach
Frage Antwort nah fern
Frage Antwort männlich weiblich
Frage Antwort warm kalt
Frage Antwort rechts links
Frage Antwort schnell langsam
Frage Antwort Junge Mädchen
[...]
Leben Tod hoch tief
Leben Tod davor danach
Leben Tod Norden Süden
Leben Tod bekannt unbekannt
Leben Tod rechts links
Leben Tod groß klein
Leben Tod warm kalt
Leben Tod männlich weiblich
Leben Tod Osten Westen
Leben Tod Sommer Winter
Mann Frau groß klein
Mann Frau schnell langsam
Mann Frau davor danach
Mann Frau lang kurz
Mann Frau oben unten
Mann Frau männlich weiblich
Mann Frau warm kalt
Mann Frau Osten Westen
Mann Frau bekannt unbekannt
Mann Frau hell dunkel
Norden Süden hell dunkel
Norden Süden gewinnen verlieren
Norden Süden groß klein
Norden Süden oben unten
Norden Süden Osten Westen
Norden Süden lang kurz
Norden Süden viel wenig
Norden Süden positiv negativ
Norden Süden Mann Frau
Norden Süden stark schwach
[...]
alt jung leicht schwer
alt jung früh spät
alt jung bekannt unbekannt
alt jung rechts links
alt jung Osten Westen
alt jung nah fern
alt jung Norden Süden
alt jung Tag Nacht
alt jung Junge Mädchen
alt jung davor danach
bekannt unbekannt leicht schwer
bekannt unbekannt rechts links
bekannt unbekannt Osten Westen
bekannt unbekannt alt jung
bekannt unbekannt schnell langsam
bekannt unbekannt Leben Tod
bekannt unbekannt viel wenig
bekannt unbekannt Mann Frau
bekannt unbekannt lachen weinen
bekannt unbekannt früh spät
[...]
leicht schwer Sommer Winter
leicht schwer warm kalt
leicht schwer hell dunkel
leicht schwer Junge Mädchen
leicht schwer Mann Frau
leicht schwer Leben Tod
leicht schwer früh spät
leicht schwer oben unten
leicht schwer lachen weinen
leicht schwer Start Ziel
männlich weiblich stark schwach
männlich weiblich Tag Nacht
männlich weiblich bekannt unbekannt
männlich weiblich lang kurz
männlich weiblich hoch tief
männlich weiblich nah fern
männlich weiblich rechts links
männlich weiblich Mann Frau
männlich weiblich Start Ziel
männlich weiblich schnell langsam
rechts links Frage Antwort
rechts links Mann Frau
rechts links hoch tief
rechts links alt jung
rechts links positiv negativ
rechts links früh spät
rechts links Start Ziel
rechts links oben unten
rechts links Junge Mädchen
rechts links lachen weinen
[...]
```

Literaturverzeichnis

[1] Kheira Leila Arras, Jan Heyd, The-Anh Ly, and Andreas Müller. Exploring semantic word similarities in German News Articles. 2014.

[2] R.E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press (Princeton, NJ), 1961.

[3] R. Bender, A. Ziegler, and St. Lange. Logistische Regression. (14):12–14, 2002.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[5] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy Layer-Wise Training of Deep Networks. *Advances in neural information processing systems*, 19(1):153, 2007.

[6] Peter F. Brown, Peter V. DeSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, (1950), 1992.

[7] Michael Curtiss, Iain Becker, and Tudor Bosman. Unicorn: a system for searching the social graph. Technical report, 2013.

[8] Joshua Goodman. A Bit of Progress in Language Modeling. page 73, 2001.

[9] Zellig Harris. *Distributional Structure*, volume 1. 1954.

[10] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982.

[11] H Kučera and W N Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.

[12] T Mikolov, M Karafiat, L Burget, J Cernocky, and S Khudanpur. Recurrent Neural Network based Language Model. *Interspeech*, (September):1045–1048, 2010.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. pages 1–9, 2013.

[14] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, 2013.

[15] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. 2012.

[16] Zach Miners. Yahoo buys SkyPhrase to better understand natural language. *PCWorld*, 2013.

[17] Marvin Minsky and Seymour Papert. *Perceptron - An Essay in Computational Geometry*. MIT Press, 1969.

[18] Frederic Morin and Y Bengio. Hierarchical probabilistic neural network language model. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252, 2005.

[19] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572, 1901.

[20] F Peng. Augmentating Naive Bayes Classifiers with Statistical Language Models. *Computer Science Department Faculty Publication Series*, Paper 91, 2003.

[21] Xin Rong. word2vec Parameter Learning Explained. pages 1–19.

[22] F Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[23] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation, 1986.

[24] David Shapiro, Doug Platts, and Magico Martinez. Google hummingbird explained. *icrossing*, 2013.

[25] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. 2013.

[26] Bernard Widrow. An Adaptive 'Adaline' Neuron Using Chemical 'Memistors', 1960.

[27] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation Classification via Convolutional Deep Neural Network. pages 1–10, 2014.