
Personalized PageRank for aligning Wikipedia articles and WordNet synsets

Bachelor-Thesis von Christian Kirschner
8. Oktober 2010



TECHNISCHE
UNIVERSITÄT
DARMSTADT



UBIQUITOUS
KNOWLEDGE
PROCESSING

Personalized PageRank for aligning Wikipedia articles and WordNet synsets

vorgelegte Bachelor-Thesis von Christian Kirschner

Supervisor: Prof. Dr. Iryna Gurevych

Coordinator: Elisabeth Wolf

Tag der Einreichung:

Erklärung zur Bachelor-Thesis

Hiermit versichere ich, die vorliegende Bachelor-Thesis ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 8. Oktober 2010

(Christian Kirschner)

Zusammenfassung

Eine Vielzahl von Anwendungen aus dem Bereich der natürlichen Sprachverarbeitung basiert auf sogenannten Wissensbasen, die Wörter und deren Bedeutungen einer oder auch mehrerer Sprachen sowie häufig die Beziehung zwischen diesen beschreiben. Beispiele für solche Wissensbasen sind WordNet, Wikipedia und Wiktionary. Eine Optimierung der Wissensbasen kann somit auch zu besseren Ergebnissen der darauf ausgeführten Anwendungen führen. Ein Ansatz ist die Verknüpfung existierender Wissensbasen zu einer erweiterten Wissensbasis, um unterschiedliche Stärken auszunutzen. Eine Möglichkeit, die Verknüpfung zu realisieren, ist die Feststellung von hinsichtlich ihrer Bedeutung übereinstimmenden Einträgen in den zu vereinigenden Wissensbasen. Dazu wird für alle möglicherweise übereinstimmenden Paare ein Ähnlichkeitswert berechnet und mit Hilfe eines trainierten Schwellenwertes entschieden, ob die Einträge eine Bedeutung teilen oder nicht.

Es existieren bereits auf einer reinen Wortüberschneidung basierende Verfahren, von denen eines als Baseline herangezogen wird. In dieser Arbeit wird eine auf dem Personalisierten PageRank Algorithmus basierende Methode vorgestellt, die auch mit semantisch ähnlichen Wörtern umgehen kann. Als Wissensbasen werden in diesem Fall WordNet und Wikipedia herangezogen, prinzipiell ist das Verfahren jedoch auf alle Wissensbasen anwendbar. Zur Evaluation wird eine entsprechende Implementierung bereitgestellt, die WordNet Synsets und Wikipedia Artikel einander zuweist. Es zeigt sich dabei, dass die vorgestellte semantische Methode in ihrer einfachen Ausführung die auf Wortüberschneidung basierende Baseline deutlich übertrifft. Insgesamt sind auf den Trainingsdaten mit denen in dieser Arbeit vorgestellten Methoden F-Measure Werte von bis zu 0.799 möglich.

Abstract

Many applications in Natural Language Processing base upon so-called lexical knowledge bases describing words and their senses of one or more languages as well as the relationship between them. WordNet, Wikipedia and Wiktionary are examples for such knowledge bases. By optimizing the knowledge bases, the results of the applications executed on them can be improved. Linking existing knowledge bases to create an extended knowledge base is one approach. This can be realized by assigning entries describing the same sense in the resources. Therefore we calculate a similarity value for each possibly matching pair. A threshold is used to classify the entries as match or not.

Some existing approaches use word overlapping methods. In this elaboration we use such a word overlapping method as baseline and present a method which is based on the Personalized PageRank algorithm and can deal with semantic similarities. We use WordNet and Wikipedia as knowledge bases, whereas in principle the procedure can be applied on all knowledge bases. For the evaluation we provide an implementation, which assigns WordNet synsets and Wikipedia articles. We will demonstrate that our semantic approach outperforms the word overlapping method (baseline) clearly. All in all we reach F-Measure values up to 0.799 on the training data with the methods proposed in this elaboration.

Inhaltsverzeichnis

1	Einleitung	5
1.1	Problemstellung und Motivation	5
1.2	Verwendete Ressourcen	5
1.2.1	WordNet	6
1.2.2	Wikipedia	6
1.3	Verknüpfung der Ressourcen	8
2	Ansätze	10
2.1	String-basierter Ansatz	10
2.2	PageRank und Personalisierter PageRank	11
2.3	Ähnlichkeitsmaße	15
2.4	Kombination der Ansätze	17
3	Umsetzung	18
3.1	Preprocessing Pipeline	18
3.2	Training/Testing Pipeline	19
4	Evaluation	24
4.1	Analyse und Bewertung der Ergebnisse	26
4.1.1	Ergebnisse ohne Bonussystem	27
4.1.2	Ergebnisse mit Bonussystem	30
4.2	Fehleranalyse	30
5	Related Work	35
6	Zusammenfassung	38
	Abbildungsverzeichnis	39
	Tabellenverzeichnis	40
	Literaturverzeichnis	41

1 Einleitung

1.1 Problemstellung und Motivation

In den vergangenen Jahren hat sich das Internet zu einer gewaltigen Ansammlung von größtenteils unstrukturierten Daten entwickelt. Das Forschungsgebiet „Natural Language Processing“ (NLP) beschäftigt sich unter anderem mit der computergestützten Erkennung der Bedeutung von Texten. Zu den wichtigsten Forschungsfeldern zählen Word Sense Disambiguation (WSD), automatische Textzusammenfassung oder maschinelle Übersetzung.

Diese verschiedenen NLP Anwendungen haben gemeinsam, dass (zumindest für sogenannte „knowledge-rich“ Ansätze) für deren Ausführung eine Wissensbasis benötigt wird. Neben einer Auflistung von Wörtern und deren möglichen Bedeutungen sind Informationen bezüglich der Beziehung zwischen den einzelnen Bedeutungen von Vorteil (sogenannte semantische Beziehungen). Die wohl am meisten verwendete englischsprachige Wissensbasis dieser Art ist „WordNet“ [Fellbaum, 1998]. Verschiedene Forschung in NLP hat jedoch gezeigt, dass auch andere Quellen wie beispielsweise Wikipedia¹ oder Wiktionary² geeignete Wissensbasen darstellen, auch wenn deutliche Unterschiede existieren [Zesch *et al.*, 2008]. So wurde insbesondere Wikipedia schon für verschiedenste NLP Anwendungen wie Textkategorisierung [Gabrilovich and Markovitch, 2006], zum Berechnen semantischer Ähnlichkeiten [Zesch *et al.*, 2007] oder für die thematisch eng mit dieser Arbeit verwandte Word Sense Disambiguation [Mihalcea, 2007] genutzt.

Um die genannten Anwendungen weiter zu verbessern, gibt es prinzipiell zwei Ansatzpunkte. Zum Einen ist weitere Forschung nach geeigneten Verfahren bzw. eine Verbesserung existierender Algorithmen notwendig. Andererseits ist eine Optimierung der Wissensbasen auf denen die Verfahren arbeiten von hoher Bedeutung. Die aus ihnen gewonnenen Informationen legen den Grundstein für eine erfolgreiche Texterkennung. In diesem Punkt liegt die Motivation für diese Arbeit, deren Ziel die Schaffung einer erweiterten Wissensbasis ist. Die Erstellung dieser erweiterten Wissensbasis soll dabei vollautomatisch durch die Verknüpfung von WordNet und Wikipedia erfolgen. Der vorgestellte Ansatz lässt sich jedoch auch ohne großen Aufwand um weitere Wissensbasen wie z.B. Wiktionary erweitern. Verschiedene Forschung auf dem Gebiet hat bereits gezeigt, dass die Verknüpfung existierender Wissensbasen zu einer Verbesserung bekannter Methoden führen kann. Ponzetto and Navigli [2010] beispielsweise zeigen, dass sich die Performanz von Word Sense Disambiguation auf einer mit Hilfe von Wikipedia erweiterten Version von WordNet verbessert. Dies ist vor allem damit begründet, dass die Ressourcen unterschiedliche Stärken besitzen, die sich durch eine Kombination besser ausnutzen lassen.

1.2 Verwendete Ressourcen

Im Folgenden soll ein Überblick über die in dieser Arbeit verwendeten Wissensbasen sowie deren Stärken und Schwächen vermittelt werden.

¹ <http://en.wikipedia.org>

² <http://en.wiktionary.org>

1.2.1 WordNet

Das semantische Wörterbuch WordNet ist die in NLP am häufigsten genutzte englischsprachige Wissensbasis. Die Bedeutungen werden durch insgesamt 117.659 sogenannte Synsets (Version 3.0) repräsentiert. Ein Synset enthält dabei die der Bedeutung zugehörigen Synonymwörter, eine kurze Beschreibung (Gloss) und teilweise einen kleinen Beispielsatz, außerdem verschiedene semantische und lexikalische Beziehungen zu anderen Synsets (Hyponym, Hyperonym, Meronym, Antonym etc.). Es gibt Synsets zu den Wortarten Nomen, Verb, Adjektiv und Adverb. WordNet wurde von Linguisten an der Princeton University entwickelt und ist frei verfügbar³.

Zu den Stärken von WordNet zählen die Berücksichtigung der vier Wortarten Nomen, Verb, Adjektiv und Adverb sowie die semantischen Relationen. Von Nachteil ist der vergleichsweise geringe Informationsgehalt: So beschränken sich die zu einem Synset angebotenen Informationen in der Regel auf ein bis zwei Sätze (Gloss). Außerdem ist die nicht optimale Abdeckung der Bedeutungen sowie die mangelnde Aktualität problematisch.

Das Nomen „actor“ beispielsweise ist zwei verschiedenen Synsets zugeordnet, hat also zwei mögliche Bedeutungen in WordNet:

- S: (n) actor, histrion, player, thespian, role player (a theatrical performer)
- S: (n) actor, doer, worker (a person who acts and gets things done) "he's a principal actor in this affair"; "when you want something done get a doer"; "he's a miracle worker"

Das erste Synset bildet unter anderem semantische Relationen (Hyperonym, Hyponym) zu den folgenden Synsets (siehe Abbildung 1.1).

Hyperonym = $\left\{ \begin{array}{l} \text{S: (n) performer, performing artist (an} \\ \text{entertainer who performs a dramatic or} \\ \text{musical work for an audience)} \end{array} \right.$

Hyponyme = $\left\{ \begin{array}{l} \text{S: (n) actress (a female actor)} \\ \text{S: (n) comedian (an actor in a comedy)} \end{array} \right.$

Mit Hilfe dieser Relationen können semantisch ähnliche Wörter und Bedeutungen bestimmt werden und Synsets genauer umschrieben werden.

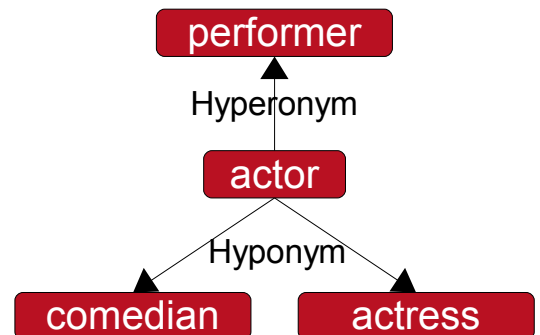


Abbildung 1.1: Semantische Relationen zwischen Synsets in WordNet

1.2.2 Wikipedia

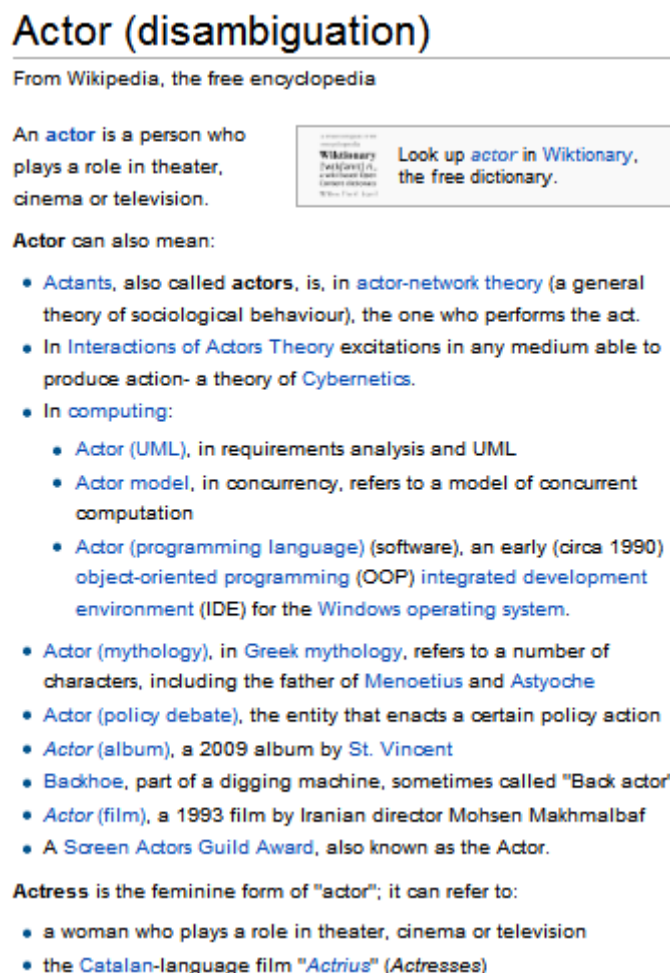
Wikipedia hat große Bekanntheit als frei verfügbare, gemeinschaftlich entwickelte Online Encyclopädie erlangt, deren Inhalte (trotz der gemeinschaftlichen Entwicklung) von hoher Qualität sind [Giles, 2005]. Von der Struktur her ergeben sich Parallelen zu gewöhnlichen Wissensbasen: Bedeutungen werden durch

³ <http://wordnet.princeton.edu/>

die einzelnen Artikel repräsentiert, Synonyme lassen sich mit sogenannten Redirects aufspüren, semantische Beziehungen erhält man über den Kategoriegraphen und die In-/Outlinks eines Artikels. Da die Daten zunächst jedoch in unstrukturierter Form vorliegen, sind entsprechende Verarbeitungsschritte notwendig, um diese semantischen Informationen nutzen zu können. Außerdem enthalten die Artikel häufig Rauschen in Form von Wörtern, die keinen Zusammenhang mit der Bedeutung des Artikels haben, sich aber kaum oder nur schwer filtern lassen (z.B. die Wörter „see terminology“ in Abbildung 1.3).

Mit seinen mittlerweile 3.362.678 englischsprachigen Artikeln (Juli 2010) erreicht Wikipedia eine deutlich höhere Abdeckung an Bedeutungen als WordNet und enthält zudem eine sehr große Menge encyclopädischen Wissens zu den einzelnen Bedeutungen. Im Unterschied zu WordNet werden auch sehr aktuelle Bedeutungen (z.B. über Filme oder Personen) abgedeckt, es werden jedoch nahezu ausschließlich Nomen beschrieben. Von großem Nutzen für viele Anwendungen ist zudem die Multilingualität von Wikipedia.

Zu dem Begriff „actor“ findet man über die Disambiguierungsseite eine ganze Reihe möglicher Artikel und somit Bedeutungen (siehe Abbildung 1.2).



Actor (disambiguation)
From Wikipedia, the free encyclopedia

An **actor** is a person who plays a role in theater, cinema or television.

Actor can also mean:

- **Actants**, also called **actors**, is, in **actor-network theory** (a general theory of sociological behaviour), the one who performs the act.
- In **Interactions of Actors Theory** excitations in any medium able to produce action- a theory of **Cybernetics**.
- In **computing**:
 - **Actor (UML)**, in requirements analysis and UML
 - **Actor model**, in concurrency, refers to a model of concurrent computation
 - **Actor (programming language)** (software), an early (circa 1990) object-oriented programming (OOP) integrated development environment (IDE) for the Windows operating system.
- **Actor (mythology)**, in **Greek mythology**, refers to a number of characters, including the father of **Menoetius** and **Astyoche**
- **Actor (policy debate)**, the entity that enacts a certain policy action
- **Actor (album)**, a 2009 album by **St. Vincent**
- **Backhoe**, part of a digging machine, sometimes called "Back actor"
- **Actor (film)**, a 1993 film by Iranian director **Mohsen Makhmalbaf**
- A **Screen Actors Guild Award**, also known as the Actor.

Actress is the feminine form of "actor"; it can refer to:

- a woman who plays a role in theater, cinema or television
- the **Catalan-language film "Actrius"** (*Actresses*)

Look up **actor** in Wiktionary, the free dictionary.

Abbildung 1.2: Beispiel einer Disambiguierungsseite von Wikipedia

Ein einzelner Artikel lässt sich in der Regel gut durch den ersten Absatz des Artikels beschreiben, welcher eine kurze Beschreibung der jeweiligen Bedeutung enthält (siehe Abbildung 1.3). Des Weiteren lässt sich die Bedeutung eines Artikels durch Kategorien und Redirects genauer beschreiben: Der Artikel „actor“ ist beispielsweise den Kategorien „Actors“, „Entertainment occupations“, „Acting und Theatrical profes-

sions“ zugeordnet und hat unter anderem die folgenden Redirects: „Theatre actor“, „Actor & Actress“, „Film actor“.

An **actor** or **actress** (see [terminology](#)) is a person who acts in a dramatic production and who works in film, television, theatre, or radio in that capacity.^[1] The ancient Greek word for an "actor," ὑποκριτής (*hypokrites*), means literally "one who interprets";^[2] in this sense, an actor is one who interprets a dramatic character.^[3]

Abbildung 1.3: Erster Absatz eines Wikipedia Artikels

1.3 Verknüpfung der Ressourcen

Wie bereits erwähnt ist das Ziel dieser Arbeit die Schaffung einer erweiterten Wissensbasis durch eine Kombination der Wissensbasen WordNet und Wikipedia. Während Wikipedia eine sehr große Menge an Informationen auch über aktuelle Themen enthält, sich allerdings auf Nomen beschränkt, sind in WordNet auch Verben, Adjektive und Adverben zu finden. Eine Kombination der beiden Ressourcen erreicht also eine deutlich höhere Abdeckung an Bedeutungen. Hinzu kommt, dass existierende WordNet Synsets durch die reichhaltigen Informationen aus Wikipedia angereichert werden können. Des Weiteren können neue semantische Beziehungen hergestellt und die Multilingualität von Wikipedia genutzt werden.

Grundvoraussetzung für die Verknüpfung der Wissensbasen ist die Feststellung und Zuweisung sinn gemäß übereinstimmender Einträge in den betrachteten Ressourcen. Der in dieser Bachelorarbeit vorgestellte Ansatz behandelt dieses Zuweisungsproblem. In diesem Fall werden einem WordNet Synset, aus einer Vorauswahl möglicher korrespondierender Wikipedia Artikel, die der jeweiligen Bedeutung entsprechenden Artikel zugewiesen. So wird dem oben betrachteten Synset „S: (n) actor, histrion, player, thespian, role player (a theatrical performer)“ der Wikipedia Artikel mit dem Namen „Actor“ zugewiesen und nicht beispielsweise der Artikel „Actor (UML)“ von der Disambiguierungsseite, welcher eine andere Bedeutung beschreibt. Zu beachten ist dabei, dass einem Synset nicht immer exakt ein Artikel zugewiesen werden kann. Es ist auch möglich, dass es keinen oder aber mehrere korrespondierende Artikel zu einem Synset gibt. So kann das Synset „S: (n) marksman, sharpshooter, crack shot (someone skilled in shooting)“ den beiden Artikeln „Marksman“ und „Sniper“ zugeordnet werden.

Für die Zuweisung muss folglich zunächst die Ähnlichkeit eines Synset-Artikel Paares berechnet werden. In Abschnitt 1.2 wurde bereits darauf hingedeutet, dass ein WordNet Synset durch Synonyme, den Gloss, sowie ein Beispiel beschrieben wird und zusätzlich semantisch ähnliche Synsets (z.B. Hyperonyme und Hyponyme) zur genaueren Eingrenzung einbezogen werden können. Ein Wikipedia Artikel wird entsprechend durch den Artikel Titel, den ersten Absatz sowie zusätzlich durch Kategorien und Redirects identifiziert. Das Mapping Problem lässt sich folglich auf den Vergleich zweier relativ kurzer, die jeweilige Bedeutung beschreibende Texte reduzieren. Welche Wörter genau in diese Beschreibung aufgenommen werden und welche Auswirkungen dies auf die Ergebnisse hat, wird die Evaluation zeigen (siehe Abschnitt 4). Vorweg ist jedoch bereits festzustellen, dass in obigem Beispiel durch die Hinzunahme entsprechender Wörter bereits eine höhere Überschneidung vorliegt: In den Redirects zum Artikel „actor“ findet sich mehrfach das Wort „actress“, welches ein Hyponym des betrachteten Synsets darstellt. Semantische Ähnlichkeit ist ebenso zwischen dem Redirect „Comic actors“ und dem Hyponym „comedian“ zu erkennen. Die Arbeit von Banerjee and Pedersen [2003] zeigt, dass die Repräsentation von WordNet Synsets durch die Einbeziehung stark verwandter Synsets verbessert werden kann. Das macht auch deshalb besonders Sinn, weil ein einzelnes Synset in der Regel nur sehr wenige verwertbare Wörter enthält, die alleine oft nicht zur Identifizierung ausreichen. Nehmen wir beispielsweise das fol-

gende „Synset S: (n) educational institution (an institution dedicated to education)“. Daraus lassen sich lediglich die vier Wörter „educational institution“, „institution“, „dedicate“ und „education“ gewinnen. Viel leichter lässt sich jedoch der korrespondierende Wikipedia Artikel „School“ zuweisen, wenn z.B. die Hyponyme hinzugefügt werden, wodurch Wörter wie „school“, „college“ und „university“ hinzukommen.

Während sich die Forschung schon intensiver mit dem Vergleich einzelner Wörter [Agirre *et al.*, 2009] oder längerer Texte (im Rahmen von Text Klassifikation und Information Retrieval) beschäftigt hat, existieren zum Vergleich kurzer Textabschnitte relativ wenige Ansätze [Mihalcea *et al.*, 2006], [Glickman *et al.*, 2005]. Der Textvergleich findet in dieser Arbeit in einem sogenannten „Bag of Words“ Ansatz statt. Bildlich gesprochen werden demnach zunächst die einzelnen für den Kontext relevanten Wörter der beiden Texte in jeweils einer Tasche zusammengefasst

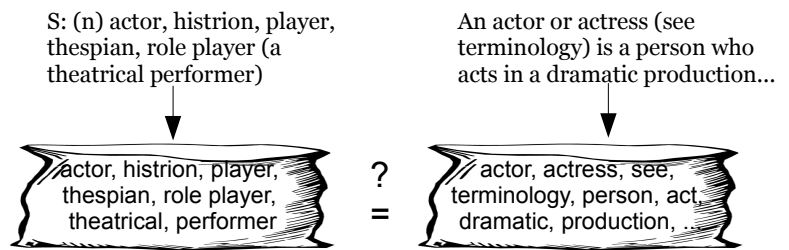


Abbildung 1.4: Erstellung zweier Bags of Words aus einem WordNet Synset (links) und dem ersten Absatz eines Wikipedia Artikels (rechts)

(siehe Abbildung 1.4). Die Reihenfolge und Anzahl an Wiederholungen der Wörter in den Texten spielt somit keine Rolle. Anschließend werden die beiden Taschen über geeignete Methoden miteinander verglichen und ein Ähnlichkeitswert bestimmt. Es existieren bereits auf einer Wortüberlappung beruhende Verfahren (siehe Kapitel 5), die jedoch keine semantischen Ähnlichkeiten berücksichtigen. Diese Arbeit beschäftigt sich mit einer auf dem Personalisierten PageRank Algorithmus basierenden Methode [Agirre and Soroa, 2009], die später im Detail vorgestellt wird. Anhand der berechneten Ähnlichkeitswerte und eines für diese Aufgabe trainierten Schwellenwertes ist dann eine Aussage darüber möglich, ob die beiden Texte inhaltlich übereinstimmen oder nicht.

2 Ansätze

Wie in Kapitel 1 erläutert findet die Zuweisung eines WordNet Synsets und eines Wikipedia Artikels auf Basis eines Bag of Words Ansatzes statt. Dazu ist es erforderlich die den zwei einander zuzuweisenden Einträgen entsprechenden Bags of Words miteinander zu vergleichen und einen Ähnlichkeitswert zu bestimmen, der dann anhand eines Schwellenwertes über eine positive bzw. negative Klassifikation entscheidet. Der hier vorgestellte Ansatz besteht, ausgehend von vorliegenden Bags of Words, aus zwei Schritten: Zunächst wird eine mathematische Repräsentation in Form eines Vektors für jede einzelne Bag of Words ermittelt. Im zweiten Schritt können diese Repräsentationen dann genutzt werden, um einen Ähnlichkeitswert zu berechnen (siehe Abbildung 2.1).

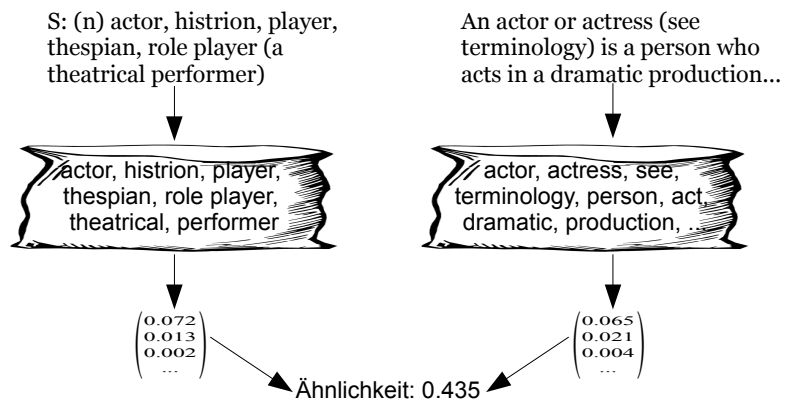


Abbildung 2.1: Die Vorgehensweise im Überblick

Im Folgenden wird zunächst ein rein String-basierter Ansatz zur Erstellung der repräsentativen Vektoren vorgestellt, der als Baseline für diese Arbeit dienen soll. Anschließend wird der (Personalisierte) PageRank Algorithmus eingeführt, sowie der darauf basierende Ansatz zur Vektorerzeugung, in dem der Schwerpunkt dieser Arbeit liegt. Im dritten Teil dieses Kapitels werden Ähnlichkeitsmaße zur Berechnung eines Ähnlichkeitswertes aus zwei Vektoren vorgestellt und bewertet, der letzte Teil beschäftigt sich mit einer Kombination des String-basierten und Personalisierten PageRank Ansatzes.

2.1 String-basierter Ansatz

Beim String-basierten Ansatz, der in dieser Arbeit als Baseline dienen soll, wird die Wortüberschneidung gemessen bzw. wieviele der vorkommenden Wörter in den beiden zu vergleichenden Bags vorzufinden sind. Dazu wird für jede der Bags zunächst eine mathematische Repräsentation in Form eines Vektors erstellt, der für jedes in einer der Bags vorkommende Wort einen Eintrag enthält. Ist das aktuell betrachtete Wort in der Bag enthalten, bekommt der Vektor für den entsprechenden Eintrag den Wert 1 zugewiesen, andernfalls den Wert 0. Abbildung 2.2 il-

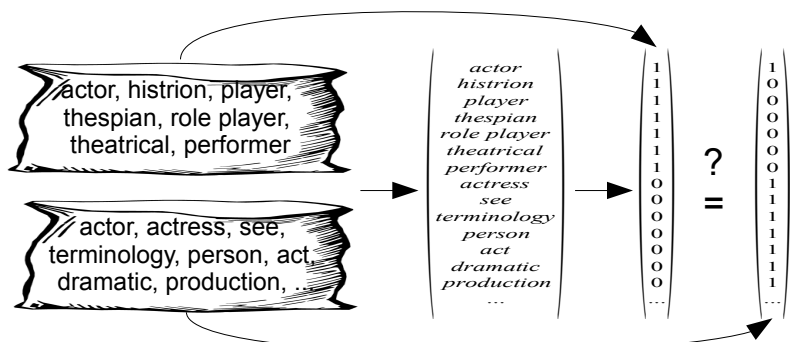


Abbildung 2.2: Erstellung von Vektoren aus Bag of Words

lustriert die Vorgehensweise. Die beiden Vektoren können mit gewöhnlichen mathematischen Methoden verglichen werden (siehe Abschnitt 2.3). Die für die Evaluation verwendete Implementation nutzt die Cosinus Distanz, die in der aktuellen Forschung sehr gute Ergebnisse erzielt. Der Nachteil dieser lediglich auf Wortüberlappung beruhenden Methode scheint offensichtlich: Semantisch ähnliche Wörter werden nicht als solche erkannt, die natürliche Sprache ist aber so vielseitig, dass der gleiche Sachverhalt auf unterschiedliche Weise ausgedrückt werden kann.

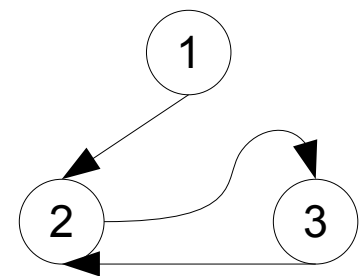
Wir betrachten beispielhaft folgendes Synset-Artikel Paar aus dem Gold-Standard:

- Wikipedia: Johannesburg also known as Jozi, Jo'burg or eGoli, is the largest city in South Africa. Johannesburg is the provincial capital of Gauteng, the wealthiest province in South Africa, having the largest economy of any metropolitan region in Sub-Saharan Africa. The city is one of the 40 largest metropolitan areas in the world, and is also the world's largest city not situated on a river, lake, or coastline.
- WordNet: city in the northeastern part of South Africa near Pretoria; commercial center for diamond and gold industries

Beide Einträge beschreiben die Stadt Johannesburg, es überschneiden sich jedoch lediglich die Wörter „city“, „South Africa“ und „Johannesburg“. Trotzdem wird jeweils die Lage und die wirtschaftliche Situation beschrieben. Dabei stehen sich beispielsweise die Wörter „large“, „economy“ (Wikipedia) und „commercial“, „center“ (WordNet) gegenüber, die eine ähnliche Bedeutung haben, aber von der String-basierten Methode als völlig verschieden angesehen werden. Dies motiviert neue Ansätze wie den im Folgenden vorgestellten PageRank-basierten Ansatz.

2.2 PageRank und Personalisierter PageRank

Der von Brin and Page [1998] vorgestellte PageRank Algorithmus ist in erster Linie von Suchmaschinen bekannt, die ihn für die Bewertung von Internetseiten verwenden. Dahinter verbirgt sich jedoch ein weitaus allgemeineres Konzept, das sich generell auf Graphstrukturen anwenden lässt. Allgemein ausgedrückt bewertet der PageRank Algorithmus die Wichtigkeit der einzelnen Knoten innerhalb eines Graphen. Jeder Knoten besitzt ein Initialgewicht, das über die Kanten im Graphen verteilt wird. Demnach erhalten Knoten mit vielen eingehenden Kanten bzw. Nachbarn von Knoten mit hohem Gewicht ebenso ein hohes Gewicht. Das Gewicht eines Knotens wird als Wahrscheinlichkeitswert angegeben. Die Summe aller Gewichte eines Graphen ergibt somit eins. Eine häufig verwendete Veranschaulichung stellt das Modell des zufälligen Läufers dar, der über den Graph läuft. Das finale Gewicht eines Knotens entspricht dabei der Wahrscheinlichkeit, dass der zufällige Läufer nach einer gewissen Zeit in diesem Knoten steht.



$$M = \begin{matrix} \text{von} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \\ \text{nach} \end{matrix} v = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$$

$$Pr = (I - cM)^{-1}(1 - c)v$$

Abbildung 2.3: PageRank

Formal ist der PageRank Wert der Knoten eines Graphen durch die folgende Gleichung bestimmt:

$$Pr = cMPr + (1 - c)v.$$

Wir nehmen einen Graphen mit n Knoten an. Zur Bestimmung der Gewichte der einzelnen Knoten muss die Gleichung nach dem n -elementigen PageRank Vektor Pr aufgelöst werden. Die $n \times n$ Matrix M (Transitionswahrscheinlichkeitsmatrix) entspricht der Wahrscheinlichkeit eines zufälligen Läufers von einem Knoten in einen Anderen zu gehen. Die Wahrscheinlichkeit ist gleichmäßig über alle ausgehenden Kanten verteilt. Der n -elementige Vektor v gibt die initiale Wahrscheinlichkeit jedes Knotens an, die im gewöhnlichen PageRank Algorithmus ebenso gleichmäßig über alle Knoten verteilt ist. Des Weiteren gibt es den sogenannten Dämpfungsfaktor c (Wertebereich $[0,1]$), der den Einfluss der beiden Summanden in der Gleichung reguliert und Konvergenz sicherstellt. Ein großer Dämpfungsfaktor führt dazu, dass das Initialgewicht stark im Graphen verteilt wird, während bei einem kleinen Dämpfungsfaktor weniger Gewicht weitergegeben wird. Abbildung 2.3 zeigt einen Beispielgraphen, die zugehörige Matrix M und den Vektor v .

Da zur Berechnung des PageRank Vektors Pr eine $n \times n$ Matrix invertiert werden muss (was bei der enormen Größe der verwendeten Graphen kaum realisierbar ist), wird der Algorithmus iterativ ausgeführt und die PageRank Werte approximiert:

$$v^{i+1} = cMv^i + (1 - c)v^0.$$

Der Vektor v^0 entspricht dem Initialvektor v , der Vektor v^k entspricht für $k \rightarrow \infty$ dem PageRank Vektor Pr . In der Regel sind jedoch 20 bis 30 Iterationen ausreichend. Abbildung 2.4 illustriert einige dieser Iterationen auf einem Beispielgraphen.

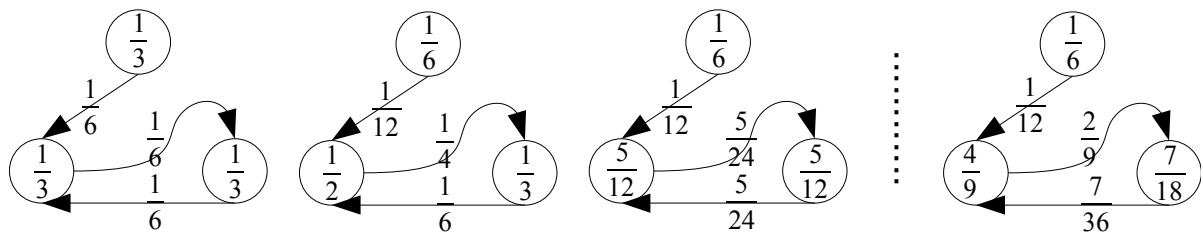


Abbildung 2.4: Einige PageRank Iterationen ($c=0.5$): Die Werte in den Knoten beschreiben die aktuellen Gewichte, die Kanten bilden die Gewichte ab, die an andere Knoten abgegeben werden. Der letzte Graph zeigt Konvergenz

Im Bereich von Natural Language Processing (NLP) zählt die Word Sense Disambiguation (WSD) zu den häufigsten Anwendungsbereichen für den PageRank Algorithmus. Das Ziel von WSD ist es, mehrdeutigen Wörtern in Texten ihre im jeweiligen Kontext korrekte Bedeutung zuzuweisen. So hat das Wort „Bank“ in den folgenden Sätzen eine unterschiedliche Bedeutung, die nur innerhalb des Kontexts deutlich wird.

- Der Mann sitzt auf einer Bank im Park.
- Die Frau geht zur Bank, um Geld abzuheben.

Um dem Wort „Bank“ seine Bedeutung zuzuordnen, sind die von Wissensbasen wie WordNet (siehe Abschnitt 1.2) angebotenen semantischen Relationen von Nutzen. So kann aus WordNet ein Graph konstruiert werden, dessen Knoten durch sämtliche WordNet Synsets repräsentiert werden, während die semantischen Relationen die Kanten bilden. Ob diese gerichtet oder ungerichtet, gewichtet oder ungewichtet sind, hängt vom jeweiligen Ansatz ab. In jedem Fall sollten in einem solchen Graphen Synsets zu den Wörtern „sitzen“ und „Park“ näher an dem korrekten Synset „Parkbank“ liegen als an dem in diesem Kontext falschen Synset „Geldinstitut“. Ebenso ist anzunehmen, dass die Bedeutungen des Wortes „Geld“

stärker mit dem Synset „Geldinstitut“ verwandt sind als mit „Parkbank“. Auf diesen Annahmen basieren aktuelle graphbasierte WSD Verfahren. Abbildung 2.5 illustriert die Idee und zeigt beispielhaft das Aussehen eines WordNet Graphen. Die Begriffe in den runden Objekten stehen für die entsprechenden Bedeutungen (Synsets). Anhand der Abbildung wird sehr schnell deutlich, wie die Kontextwörter über die korrekte Bedeutung des Wortes „Bank“ entscheiden: Sie sammeln sich in direkter Umgebung des Wortes.

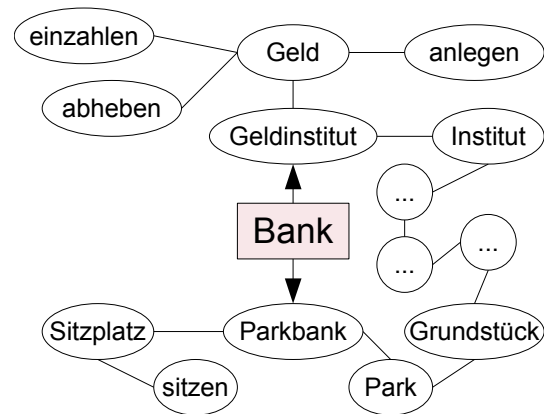


Abbildung 2.5: Beispiel für WordNet Graph mit zwei verschiedenen Bedeutungen für das Wort „Bank“

Wendet man den PageRank Algorithmus auf dem aus einer Wissensbasis konstruierten Graphen an, so erhält man folglich eine kontextunabhängige Bewertung der Wichtigkeit von Bedeutungen, was nicht zielführend ist, da die Wichtigkeit von möglichen Bedeutungen eines Wortes in Abhängigkeit des Kontextes bewertet werden soll. Um den Kontext einzubringen, gibt es im Wesentlichen zwei Möglichkeiten: Der „traditionelle“ Ansatz sieht vor, aus dem WordNet Graphen in Abhängigkeit des Kontexts einen Subgraphen zu extrahieren und darauf den PageRank Algorithmus anzuwenden. Für die Subgraphgenerierung existieren zahlreiche Variationen ([Agirre and Soroa, 2008], [Navigli and Lapata, 2007], [Mihalcea, 2005]). In jedem Fall enthält der Subgraph alle möglichen Bedeutungen der Kontextwörter, sowie in der Regel die sie auf dem kürzesten Weg verbindenden Knoten und Kanten. Alle Knoten in diesem Subgraph enthalten entsprechend des gewöhnlichen PageRank Verfahrens das gleiche Initialgewicht, der resultierende PageRank Vektor enthält dann die Werte für alle im Subgraphen enthaltenen Knoten.

Eine zweite Möglichkeit den Kontext einzubringen, wurde von Agirre and Soroa [2009] vorgestellt und sieht eine „Personalisierung“ des PageRank Algorithmus vor. Der Vorteil dieser Methode ist, dass der Algorithmus auf dem gesamten Graphen ausgeführt werden kann. Dies bedeutet einen erheblichen Effizienzgewinn, da die relativ aufwendige Subgraphextraktion damit hinfällig wird. Der Unterschied zu dem oben vorgestellten gewöhnlichen PageRank Algorithmus liegt in dem Initialisierungsvektor v . Die Gewichte in diesem Vektor sind bei der Personalisierten PageRank Methode nicht gleichmäßig auf alle Knoten des Graphen verteilt, sondern auf alle möglichen Bedeutungen der Kontextwörter. Über diese wird das Gewicht dann im Graphen verteilt. Die Disambiguierung der Kontextwörter findet (wie auch bei der Subgraphmethode) statt, indem jedem Kontextwort aus den möglichen Bedeutungen diejenige mit dem höchsten PageRank Wert zugewiesen wird. Abbildung 2.6 illustriert die beiden Vorgehensweisen: Während bei der Subgraphmethode aus dem Kontext ein Subgraph generiert wird und dann das Gewicht gleichmäßig verteilt wird, erhalten beim Personalisierten PageRank Verfahren ausschließlich die Bedeutungen der Kontextwörter das Initialgewicht. Die Idee, auf der das PageRank Verfahren beruht, ist die, dass sich das Gewicht der im Kontext falschen Bedeutungen in den Weiten des Graphen verläuft, da diese eher isoliert im Graphen liegen, während die korrekten Bedeutungen das Gewicht stärker konzentrieren, da sie dicht beieinander liegen und sich so gegenseitig Gewicht abgeben. Aus diesem Grund spricht man beim PageRank Algorithmus auch von einem Centrality oder Connectivity Measure.

Für unseren semantischen Ansatz zum Vergleich kurzer Texte auf Basis eines Bag of Words Ansatzes verwenden wir den Personalisierten PageRank Algorithmus von Agirre and Soroa [2009]. Genau wie bei WSD wird das Initialgewicht gleichmäßig über alle möglichen Bedeutungen der Kontextwörter verteilt, wobei der Kontext in diesem Fall durch eine Bag of Words repräsentiert wird. Der daraus resultierende PageRank Vektor stellt dann, ähnlich wie beim String-basierten Ansatz (siehe Abschnitt 2.1), eine

mathematische Repräsentation der entsprechenden Bag of Words dar. Abbildung 2.7 illustriert die Vorgehensweise. Dieser Ansatz kann als semantisch bezeichnet werden, da semantisch ähnliche Wörter im Graphen nahe beieinander liegen und somit durch die Weitergabe von Gewichten berücksichtigt werden.

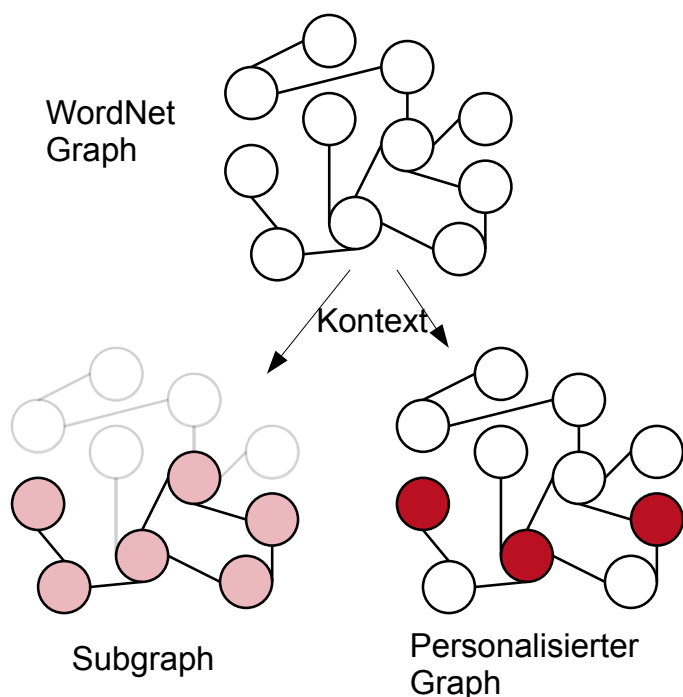


Abbildung 2.6: Vorgehensweise von Subgraph- und Personalisierter PageRank Methode (bei WSD)

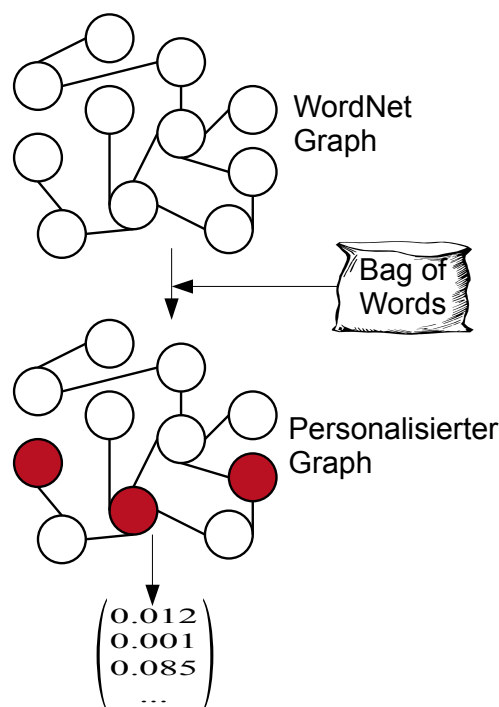


Abbildung 2.7: Erstellung repräsentativer Vektoren mit Personalisiertem PageRank

Neben den Effizienzvorteilen gegenüber der Subgraphmethode hat der Personalisierte PageRank Ansatz den Vorteil, dass die berechneten PageRank Vektoren unabhängig von den zu gewichtenden Kontextwörtern den gleichen Aufbau haben: Da der Algorithmus immer auf dem gesamten Graphen ausgeführt wird, enthalten alle daraus resultierenden PageRank Vektoren Werte für sämtliche Knoten des Graphen und sind somit unabhängig von dem Synset-Artikel Paar, für das sie erstellt werden. Würden wir eine Subgraphmethode verwenden, müssten wir den Subgraphen und somit den PageRank Vektor in Abhängigkeit des jeweils betrachteten Paares erstellen, denn wir können zwei Vektoren nur miteinander vergleichen, wenn sie den gleichen Aufbau besitzen und Werte für die gleichen Bedeutungen enthalten. Bei der oben eingeführten String-basierten Methode werden die Vektoren ebenfalls für jedes Paar erstellt. Dadurch müssen insgesamt $2 \cdot |Paare|$ Vektoren erstellt werden, während bei Verwendung der Personalisierten PageRank Methode $|Synsets| + |Artikel|$ Vektoren berechnet werden müssen. In den für die Evaluation genutzten Trainingsdaten wird ein Synset im Durchschnitt mit fast 6 Artikeln verglichen. Demnach müssten bei der Subgraphmethode (oder beim String-basierten Ansatz) für ein Synset und 6 Wikipedia Artikel 12 Vektoren erstellt werden, während der Personalisierte PageRank Ansatz mit maximal 7 Vektoren auskommt. Daraus folgt ein deutlich geringerer Aufwand bei der Nutzung des Personalisierten PageRank Algorithmus im Vergleich zur Nutzung einer Subgraphmethode.

Ähnlich zu den meisten WSD Anwendungen bauen wir den Graphen für unseren Ansatz wie schon oben beschrieben aus WordNet auf: Die Knoten werden durch Synsets, die Kanten (ungerichtet und ungewichtet) durch semantische Beziehungen repräsentiert. WordNet eignet sich von den verfügbaren Wissensbasen am ehesten, da es die wichtigsten Bedeutungen sämtlicher Wortarten und relativ gute semantische Beziehungen enthält, die mit entsprechenden Methoden noch zusätzlich erweitert werden können. So kann der Graph mit eXtended WordNet [Mihalcea and Moldovan, 2001] mit zusätzlichen

semantischen Relationen angereichert werden. Dazu werden die Gloss-Wörter der Synsets disambiguiert und zusätzliche Relationen zwischen den Synsets und den disambiguierten Bedeutungen geschaffen. Ein weiterer Vorteil, den Graphen aus WordNet aufzubauen, ist der, dass uns die Bedeutungen der WordNet Synsets, die wir mit Wikipedia Artikeln vergleichen wollen, bekannt sind. Anstatt also den Bedeutungen der Wörter, die das Synset beschreiben, Gewicht zuzuweisen, können wir auch auf die Erstellung der Bags of Words verzichten und den bekannten WordNet Synset initial das volle Gewicht zuteilen. Die Evaluation wird zeigen, wie gut dies im Vergleich zur gewöhnlichen Methode funktioniert. Wir halten uns sowohl bei den genutzten Graphen als auch bei dem Personalisierten PageRank Algorithmus an Agirre and Soroa [2009], deren Implementation wir auch für die Berechnung der für den hier vorgestellten Ansatz benötigten Vektoren nutzen.

Neben dem PageRank Algorithmus existieren weitere sogenannte Centrality oder Connectivity Measures, deren Ziel ebenfalls die Bewertung der Wichtigkeit von Knoten innerhalb eines Graphen ist. Bekannte Beispiele hierfür sind Degree, Betweenness oder Maximum Flow [Navigli and Lapata, 2007]. Wir haben uns in dieser Arbeit für den PageRank Algorithmus entschieden, weil der Algorithmus in der aktuellen Forschung weit verbreitet ist, damit in WSD bereits sehr gute Ergebnisse erzielt wurden und es zudem eine sehr gute, frei verfügbare Implementierung¹ gibt, die wir nutzen.

2.3 Ähnlichkeitsmaße

Wurden für zwei zu vergleichende Bags of Words mittels Personalisiertem PageRank (oder der String-basierten Methode) mathematische Repräsentationen in Form zweier Vektoren berechnet, so ist es möglich diese Vektoren mit gewöhnlichen mathematischen Methoden zu vergleichen und einen Ähnlichkeitswert zu bestimmen. Die Wertebereiche der Maße beziehen sich auf die für den PageRank-basierten Ansatz geltende Annahme, dass für die Vektoren A und B gilt: $\sum_i A_i = \sum_i B_i = 1$.

Maß	Formel	Intervall
Cosinus Distanz	$cosDist(A, B) = \cos(\alpha) = \frac{\sum_i A_i \cdot B_i}{\sqrt{\sum_i A_i^2 \cdot \sum_i B_i^2}}$	[0, 1]
Intersection	$int(A, B) = \sum_i \min(A_i, B_i)$	[0, 1]
Chi-square*	$\chi^2(A, B) = \sum_i \frac{(A_i - B_i)^2}{A_i + B_i}$	[0, ∞]
Euklidische Distanz	$euclid(A, B) = \sqrt{\sum_i (A_i - B_i)^2}$	[0, $\sqrt{2}$]
Punktprodukt	$dot(A, B) = \sum_i A_i \cdot B_i$	[0, 1]

* wobei in der Summe Einträge i mit $A_i + B_i < \epsilon$ ignoriert werden

Von den hier vorgestellten Ähnlichkeitsmaßen für Vektoren gehört insbesondere die Cosinus Distanz zum Standard in vielen entsprechenden NLP Anwendungen. Diese beschreibt den Winkel zwischen zwei Vektoren (siehe Abbildung 2.8). Es ist jedoch nicht möglich, eine allgemeine Aussage darüber zu treffen, welches der Maße zu den besten Ergebnissen führt, da dies sehr stark von der jeweiligen Anwendung und dem verwendeten Ansatz abhängt. Im Folgenden sollen daher vor allem die Unterschiede herausgearbeitet werden. Zu beachten ist, dass die Maße Cosinus Distanz, Intersection und Punktprodukt maximiert, während χ^2 und Euklidische Distanz minimiert werden müssen.

¹ <http://ixa2.si.ehu.es/ukb/>

Das Intersection Maß misst den gemeinsamen Teil zweier Vektoren, während sich die Euklidische Distanz auf die Unterschiede konzentriert. Dabei ist zu beachten, dass die Euklidische Distanz die einzelnen Differenzen quadriert, wodurch kleinere Distanzen weniger Einfluss haben als Größere. In beiden Fällen werden alle Einträge gleich gewichtet, was die Maße nicht besonders diskriminant macht. Anders in dieser Hinsicht ist das Maß χ^2 , bei dem eine Gewichtung statt findet: Die Distanz der Werte 0,1 und 0,2 in χ^2 ist äquivalent zur Distanz von 0,5 und 0,7 während alle anderen Maße die Distanz zwischen den ersten beiden Werten als kleiner bewerten. Das heißt für zwei relativ große Werte mit einer Differenz von x ergibt sich ein höherer Ähnlichkeitswert, als bei zwei relativ kleinen Werten mit der gleichen Differenz x . Anders ausgedrückt dürfen große Werte eine größere Differenz aufweisen als Kleinere. Auf diese Weise dominieren einzelne größere Abweichungen zwischen generell größeren Werten den berechneten Ähnlichkeitswert nicht so stark, wie es sonst der Fall wäre. Bei χ^2 muss zudem darauf geachtet werden, dass die Summe der aktuell betrachteten Werte nicht 0 ergibt, was zu einer Division durch 0 führen würde. Es ist ratsam dabei nicht nur Werte in der Berechnung zu ignorieren deren Summe exakt 0 ergibt, sondern auch Werte auszuschließen deren Summe kleiner einem Epsilon Wert ist. Dies macht bei den enorm großen PageRank Vektoren Sinn, da hier sehr viele Einträge mit sehr kleinen Werten vorkommen, die für die Berechnung weniger relevant sind, durch die Gewichtung jedoch einen zu hohen Einfluss erhalten würden. Der Epsilon Wert wurde nach einigen Tests mit aus dem Personalisierten PageRank Verfahren berechneten PageRank Vektoren auf 10^{-4} festgesetzt, ist jedoch nicht optimiert.

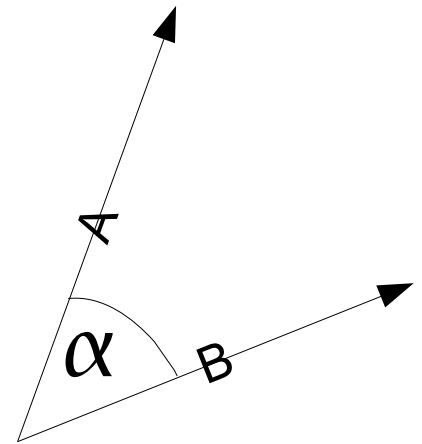


Abbildung 2.8: Cosinus Distanz

Da bei dem String-basierten Ansatz (siehe Abschnitt 2.1), der ausschließlich die Cosinus Distanz verwendet, in den Vektoren nur die Werte 0 und 1 vorkommen, ist eine Gewichtung in diesem Fall weniger sinnvoll. Hier würde χ^2 genau wie die Euklidische Distanz lediglich die Anzahl der nicht übereinstimmenden Einträge in den Vektoren zählen, sodass mit der Cosinus Distanz als Standardmaß wohl bessere Ergebnisse erzielt werden können. Auch für die anderen Maße sind beim String-basierten Ansatz keine Verbesserungen zu erwarten: Intersection und Punktprodukt zählen die Anzahl der Einträge bei denen beide Werte gleich 1 sind. Für die drei Vektorpaare A, B und C aus Abbildung 2.9 lassen sich die Paare mit der Cosinus Distanz aufsteigend nach Ähnlichkeit in der Reihenfolge A, B, C sortieren. Für Intersection und Punktprodukt ergibt sich die Reihenfolge A/B, C (A und B erhalten den gleichen Ähnlichkeitswert), für χ^2 und Euklidische Distanz A, B/C (B und C erhalten den gleichen Ähnlichkeitswert). Die Cosinus Distanz erweist sich in diesem Beispiel folglich als diskriminanter. Falls man dennoch weitere Ähnlichkeitsmaße für den String-basierten Ansatz verwenden möchte, sollte beachtet werden, dass eine Normalisierung der Ähnlichkeitswerte erforderlich ist. Bei den PageRank Vektoren ist diese nicht notwendig, da die Summe über alle Einträge eines PageRank Vektors bereits mit 1 normalisiert ist.

$$A: \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad B: \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad C: \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Abbildung 2.9: Beispiel von zu vergleichenden Vektoren (String-basierter Ansatz)

Das Punktprodukt hat die besondere Eigenschaft, dass beispielsweise der Vektor $[0.5, 0.5, 0]$ mit sich selbst verglichen einen Wert von 0.5 ergibt, während der Vektor $[1, 0, 0]$ den Maximalwert 1 mit sich selbst erreicht. Alle anderen Maße ermitteln für zwei identische Vektoren grundsätzlich die Maximalübereinstimmung. Somit werden Vektoren, die ihre Werte sehr stark konzentrieren, bevorzugt. Die Evaluation (siehe Abschnitt 4) wird zeigen, dass sich dies auf die Ergebnisse der hier besprochenen Anwendung eher negativ auswirkt. Festzustellen ist auch, dass sich die Cosinus Distanz und das Punktprodukt lediglich in dem bei der Cosinus Distanz durchgeführten Normalisierungsschritt unterscheiden, der aber scheinbar von hoher Bedeutung ist. Ebenso ist eine gewisse Ähnlichkeit zwischen χ^2 und Euklidischer Distanz erkennbar, wobei die bei χ^2 vorhandene Gewichtung mit $A_i + B_i$ bei der Euklidischen Distanz fehlt.

2.4 Kombination der Ansätze

Motiviert durch die Evaluation, die ergab, dass die beiden soeben vorgestellten Ansätze sich in den gemachten Fehlern nicht unerheblich unterscheiden, sowie relativ hohe Recall Werte² der betrachteten Ansätze, haben wir uns dazu entschlossen, eine einfache Kombination der Ansätze zu versuchen: Ein Synset-Artikel Paar wird demnach nur dann als positiv (übereinstimmend) klassifiziert, wenn sowohl der String-basierte als auch der PageRank-basierte Ansatz eine positive Klassifikation vornehmen. In allen anderen Fällen wird das Paar negativ klassifiziert.

² recall = True Positives / (True Positives + False Negatives)

3 Umsetzung

Im Folgenden soll konkret auf die Umsetzung der in Kapitel 2 vorgestellten Ansätze zum Vergleich kurzer Textabschnitte eingegangen werden. Die Implementierung befasst sich mit der Zuweisung von WordNet Synsets und Wikipedia Artikeln, ist jedoch prinzipiell ohne großen Aufwand um weitere Wissensbasen wie beispielsweise Wiktionary erweiterbar.

Umgesetzt wurde der Ansatz in der Programmiersprache Java¹ basierend auf dem Apache UIMA Projekt². Dieses Framework für Textannotationen ermöglicht auf sehr komfortable Weise die Erstellung von einzelnen unabhängigen Annotatoren, die in sogenannten Pipelines hintereinander geschaltet werden können. Dadurch ist eine gute Austauschbarkeit der einzelnen Bestandteile und eine hohe Flexibilität der gesamten Umsetzung gegeben.

Die Implementierung besteht aus einer Preprocessing Pipeline, die für die Erstellung der Bags of Words zuständig ist, sowie einer Training/Testing Pipeline, welche die Ähnlichkeiten für alle gewünschten Paare berechnet und auf den Trainingsdaten einen Schwellenwert für eine positive bzw. negative Zuweisung von Synset und Artikel lernt. Der gelernte Schwellenwert kann anschließend auf einem Testdatensatz getestet werden. Außerdem ist eine Evaluierung mittels Cross-Validation möglich. In den folgenden Abschnitten werden die beiden Pipelines im Detail betrachtet.

3.1 Preprocessing Pipeline

Der erste Schritt bei der Zuweisung von WordNet Synsets und Wikipedia Artikeln besteht darin, die Bags of Words zu erstellen. Abbildung 3.1 illustriert den Aufbau und Ablauf der entsprechenden Pipeline.

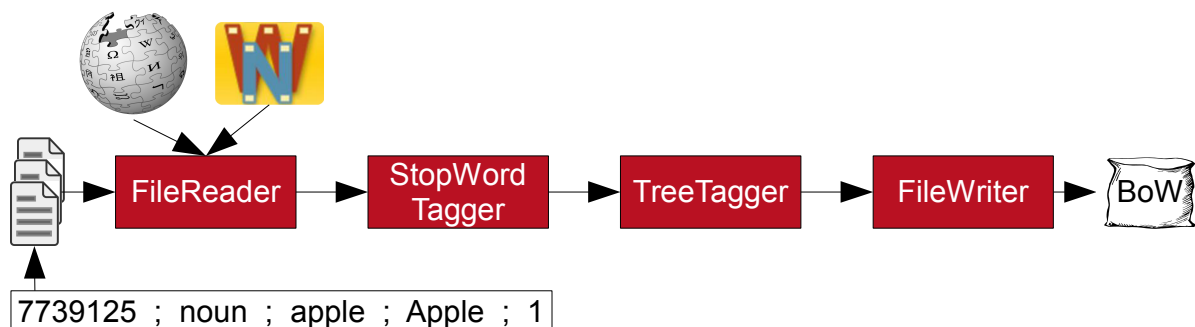


Abbildung 3.1: Preprocessing Pipeline

Zunächst wird vom „**FileReader**“ der Gold-Standard eingelesen (siehe Abbildung 3.2), der sämtliche Paare enthält, deren Ähnlichkeit bestimmt werden soll. Der Text wird von einem Parser durchlaufen, der alle Synsets und Artikel extrahiert und in eine Liste einfügt. Es spielt dabei keine Rolle in welchen Paarungen die Synsets bzw. Artikel vorkommen, zu jedem im Gold-Standard einfach oder mehrfach

¹ <http://www.java.com/de/>

² <http://uima.apache.org/>

vorkommenden Synset bzw. Artikel muss nur eine Bag of Words erstellt werden.

```
# Each row represents a sense pair consisting of a WordNet 3.0 synset and a Wikipedia article (dump version: August, 22nd 2009)
# annotated with "1" (same sense) or "0" (different sense)
#
# Synset offset ; POS ; Synset ; Wikipedia article title ; Annotation
#
5077146 ; noun ; alignment ; Alignment (role-playing games) ; 0
5077146 ; noun ; alignment ; Alignment (Dungeons & Dragons) ; 0
5077146 ; noun ; alignment ; Alignment (political party) ; 0
5077146 ; noun ; alignment ; Alignment (archaeology) ; 0
```

Abbildung 3.2: Ausschnitt aus dem Gold-Standard

Dazu wird die Liste mit den extrahierten Synsets bzw. Artikeln durchlaufen und nacheinander für jeden einzelnen Eintrag der folgende Arbeitsablauf ausgeführt:

Zunächst wird aus der jeweiligen API das zugehörige WordNet Synset bzw. der zugehörige Wikipedia Artikel ausgelesen. Die Implementation ermöglicht dabei eine genaue Auswahl der in eine Bag of Words aufgenommenen Wörter: So ist es möglich neben dem Synset selbst Hyperonyme und/oder Hyponyme mit aufzunehmen bzw. neben dem ersten Absatz eines Artikels und dem Artikel Titel auch Redirects und/oder Kategorien. Für den Zugriff auf WordNet wird in dieser Arbeit die Java WordNet Library (JWNL)³ genutzt, auf Wikipedia wird über die Java Wikipedia Library (JWPL)⁴ zugegriffen. Aus mehreren Einzelwörtern zusammengesetzte Wörter (z.B. United States of America) werden vor weiteren Verarbeitungsschritten zusammengefasst indem für jeweils bis zu vier aufeinander folgende Wörter geprüft wird, ob sie als ein Begriff in WordNet eingetragen sind.

Die aus den APIs ausgelesenen Texte liegen größtenteils in natürlicher Sprache vor. Diese enthält viele Füllwörter, die für den Kontext ohne Bedeutung sind (z.B. Artikel oder Hilfsverben). Der „**StopWordTagger**“ hat die Aufgabe solche sogenannten StopWords aus den Texten zu entfernen. Um die verbleibenden Wörter vergleichen zu können und um die den Wörtern entsprechenden Bedeutungen im WordNet Graphen für den Personalisierten PageRank Algorithmus zu finden, ist es zudem notwendig sie in ihre Grundform zu überführen und die Wortart zu bestimmen. Diese Aufgabe wird vom „**TreeTagger**“⁵, einem frei verfügbaren Lemmatisierer, übernommen. Abschließend werden die Wörter in ihrer Grundform und mit ihrer jeweiligen Wortart (Nomen, Verb, Adjektiv, Adverb) vom „**FileWriter**“ in eine Textdatei geschrieben, die somit eine Bag of Words für jeweils ein WordNet Synset bzw. einen Wikipedia Artikel darstellt. Abbildung 3.3 zeigt beispielhaft einen Ausschnitt für die resultierende Bag of Words für den Wikipedia Artikel „Actor“.

```
Actor
actor#n#w1#1 actress#n#w2#1 see#v#w3#1
terminology#n#w4#1 person#n#w5#1 act#v#w6#1
dramatic#a#w7#1 production#n#w8#1 ...
```

Abbildung 3.3: Bag of Words zum Wikipedia Artikel „Actor“

3.2 Training/Testing Pipeline

Die zweite Pipeline (siehe Abbildung 3.4) initiiert die Berechnung von PageRank Vektoren für zuvor mit der Preprocessing Pipeline erstellte Bags of Words, berechnet die Ähnlichkeit von Synset-Artikel

³ <http://sourceforge.net/projects/jwordnet/>

⁴ <http://www.ukp.tu-darmstadt.de/software/jwpl/>

⁵ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

Paaren, trainiert einen Schwellenwert und klassifiziert eine gegebene Menge an Paaren hinsichtlich ihrer Bedeutung als gleich oder verschieden. Die Annotatoren können je nach Wunsch so zusammengesetzt werden, dass es möglich ist die Pipeline zum Trainieren auf dem Gold-Standard, zum Testen auf einem Testdatensatz, zum Evaluieren mittels Cross-Validation oder zum einfachen Klassifizieren zu verwenden. Benötigt werden dazu lediglich die mit der Preprocessing Pipeline berechneten Bags of Words in Form von Textdateien.

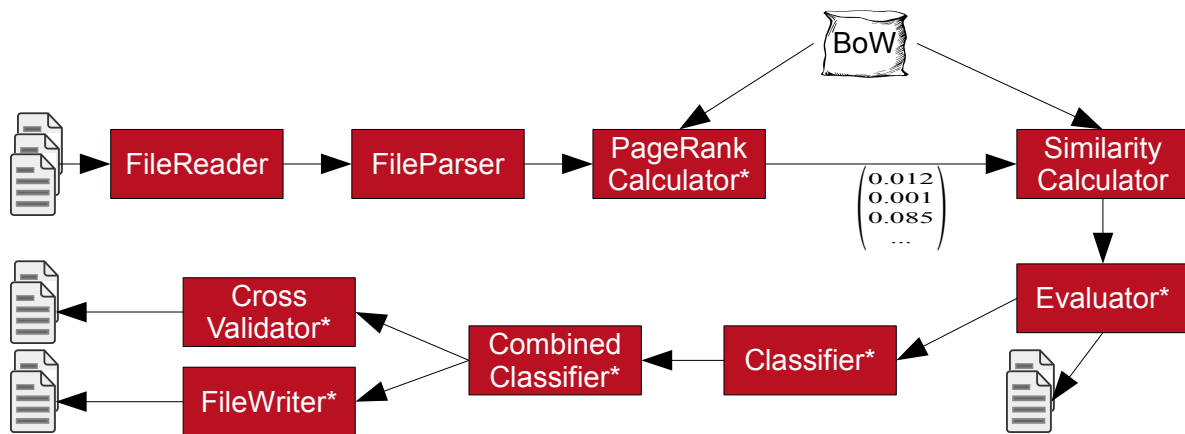


Abbildung 3.4: Training/Testing Pipeline (* Annotator ist je nach Aufbau der Pipeline optional)

Die Arbeit der Pipeline beginnt grundsätzlich mit dem Einlesen des Trainings-/Testdatensatzes (Gold-Standard) mit dem „**FileReader**“. Der „**FileParser**“ durchläuft den Datensatz und speichert wie schon in der Preprocessing Pipeline die einzelnen Synsets und Wikipedia Artikel in Listen ab. Zusätzlich werden noch die zu vergleichenden Paare, sowie die jeweilige Annotation (sofern vorhanden) für den weiteren Ablauf gesichert. Falls die Pipeline zum Berechnen von Ähnlichkeiten mit der Personalisierten PageRank Methode genutzt werden soll, ist im Folgenden der „**PageRankCalculator**“ notwendig. Dieser berechnet mit der Methode des Personalisierten PageRank Algorithmus (siehe Abschnitt 2.2) für alle Synsets bzw. Artikel der Liste aus den zuvor mit der Preprocessing Pipeline erstellten Bags of Words die zugehörigen PageRank Vektoren und speichert diese in Textdateien. Dies kann je nach Größe des Datensatzes und verfügbarer Prozessorkapazität viel Zeit in Anspruch nehmen. Bereits berechnete PageRank Vektoren brauchen daher bei einem späteren Durchlauf nicht erneut berechnet werden. Für die Berechnung der PageRank Vektoren wird die in Agirre and Soroa [2009] für Word Sense Disambiguation genutzte Implementierung des Personalisierten PageRank Verfahrens genutzt⁶. Der PageRank Algorithmus kann wahlweise auf einem WordNet Graphen der Version 1.7, angereichert mit semantischen Relationen durch eXtended WordNet [Mihalcea and Moldovan, 2001] oder der Version 3.0, angereichert mit manuell disambiguierten Relationen, ausgeführt werden. Sofern nicht die Personalisierte PageRank Methode angewandt werden soll (sondern der String-basierte Ansatz), kann der „**PageRankCalculator**“ weggelassen werden, da für die weiteren Berechnungen dann keine PageRank Vektoren benötigt werden.

Der „**SimilarityCalculator**“ berechnet im Anschluss Ähnlichkeitswerte für alle Paare mit den gewünschten Ähnlichkeitsmaßen. Zur Auswahl stehen neben den in Abschnitt 2.3 vorgestellten Maßen für den PageRank basierten Ansatz (Cosinus Distanz, χ^2 , Intersection, Euklidische Distanz, Punktprodukt) die rein String-basierte Methode (siehe Abschnitt 2.1) und eine rein zufalls-basierte Methode, die jedem Paar einen zufälligen Ähnlichkeitswert zwischen 0 und 1 zuweist und neben der String-basierten Methode als Baseline herangezogen werden soll. Es ist dabei möglich alle Methoden parallel zu berechnen. Für die String-basierte Methode sind dabei die Bags of Words, jedoch keine PageRank Vektoren, notwendig.

⁶ <http://ixa2.si.ehu.es/ukb/>

Falls auf einem Trainingsdatensatz gearbeitet wird, kann der „**Evaluator**“ dazu genutzt werden mit Hilfe der berechneten Ähnlichkeiten auf dem Datensatz einen optimalen Schwellenwert für eine Zuweisung zu trainieren. Dies geschieht, indem alle zuvor berechneten Ähnlichkeitswerte und zusätzlich Werte in $\frac{1}{100}$ Schritten als mögliche Schwellenwerte angesehen werden, zu denen Auswertungsmaße wie Genauigkeit und F-Measure berechnet werden. Es wird dann der Schwellenwert gewählt, der das gewünschte Maß (in dem Fall F-Measure) optimiert. Falls mehrere Schwellenwerte mit maximalem F-Measure gefunden werden, wird der kleinste gewählt. Mehr zu den Auswertungsmaßen folgt in Kapitel 4. Alternativ ist es möglich für eine 10-fold Cross-Validation Schwellenwerte für die einzelnen Folds zu ermitteln. Dazu wird der Datensatz in 10 möglichst gleich große Folds unterteilt. Dann werden jeweils 9 der 10 Folds genutzt, um einen Schwellenwert zu trainieren, woraus sich insgesamt 10 Schwellenwerte ergeben. Neben dem Training gibt der „Evaluator“ eine Reihe von Statistiken und Graphen wie Recall-Precision Kurve, ROC-Kurve etc. aus (siehe Abbildungen 3.5 und 3.6), die Aufschluss über die Performanz geben.

Soll die Pipeline nur zur Klassifikation oder zum Testen mit festem, manuell eingegebenen Schwellenwert genutzt werden, kann der „Evaluator“ weggelassen werden. Für eine Cross-Validation ist er jedoch in jedem Fall erforderlich. Falls der Nutzer lediglich die Absicht hat einen Schwellenwert zu trainieren und keine Klassifikation durchzuführen, können alle weiteren Elemente der Pipeline weggelassen werden. Andernfalls folgt alternativlos der „Classifier“.

Der „**Classifier**“ ist für die Klassifikation von Paaren zuständig (positiv = übereinstimmend, negativ = verschieden). Der Annotator klassifiziert die Paare anhand der zuvor vom „SimilarityCalculator“ berechneten Ähnlichkeitswerte, sowie eines vom „Evaluator“ übergebenen oder manuell eingegebenen Schwellenwertes. Falls eine Cross-Validation durchgeführt werden soll, werden die 10 Folds jeweils mit dem Schwellenwert klassifiziert, bei dessen Training das entsprechende Fold ausgelassen wurde. Der „**CombinedClassifier**“ führt zusätzlich noch eine Klassifikation für die Kombinationsmethode durch bei der ein Paar nur dann als positiv klassifiziert wird, wenn sowohl die String-basierte Methode als auch die PageRank-basierte Methode das Paar positiv klassifizieren (siehe Abschnitt 2.4). Dieser Annotator macht nur Sinn, wenn im „SimilarityAnnotator“ mehrere Methoden zur Berechnung eingegeben wurden. Andernfalls kann er weggelassen werden.

Sofern man die Pipeline für eine Cross-Validation nutzt, dient der „**CrossValidator**“ der Ausgabe entsprechender Auswertungsdaten in Textdateien (siehe Abbildung 3.7). Zu jedem der 10 Folds werden Daten wie F-Measure, Accuracy oder der gewählte Schwellenwert gespeichert. Zusätzlich ist den Dateien jeweils ein Durchschnittswert über die Folds, sowie die Standardabweichung zu entnehmen. Des Weiteren gibt der „**FileWriter**“ die Klassifikation sämtlicher Paare, sowie einige Auswertungsdaten in Textdateien aus. Die Reihenfolge von „CrossValidator“ und „FileWriter“ ist dabei beliebig.

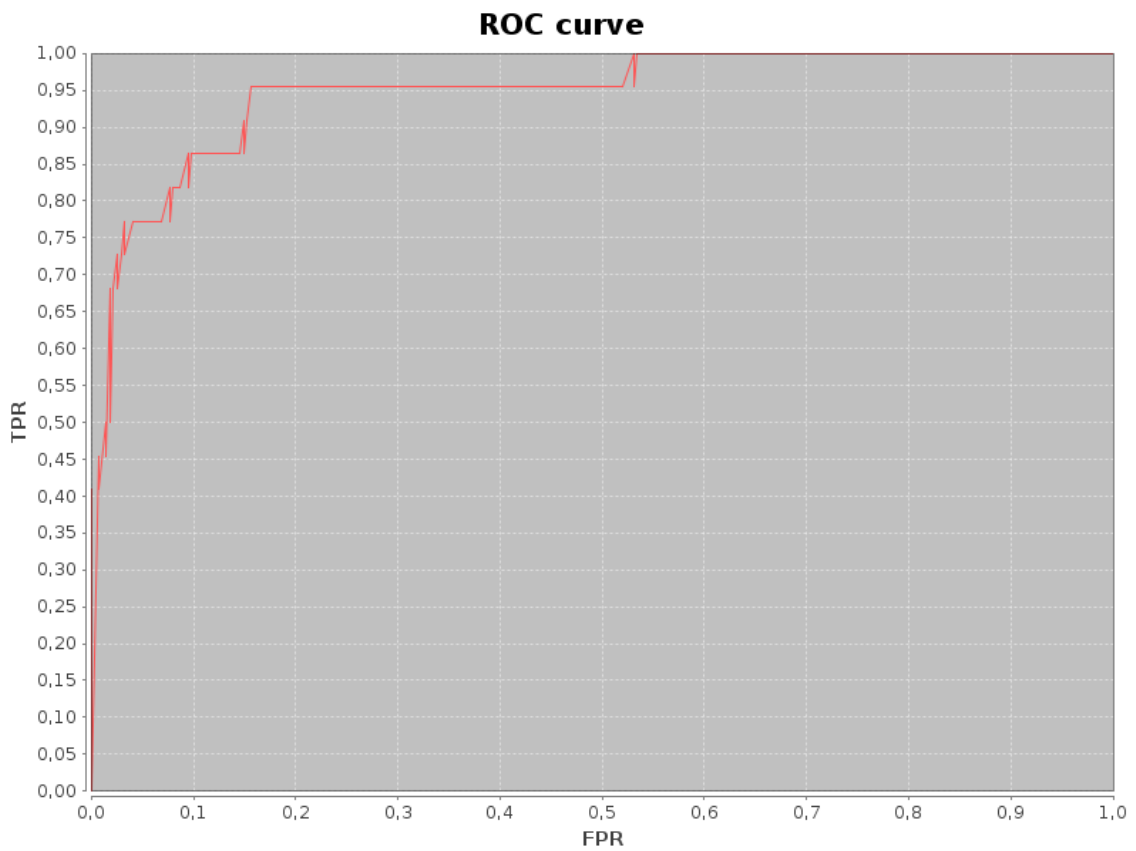
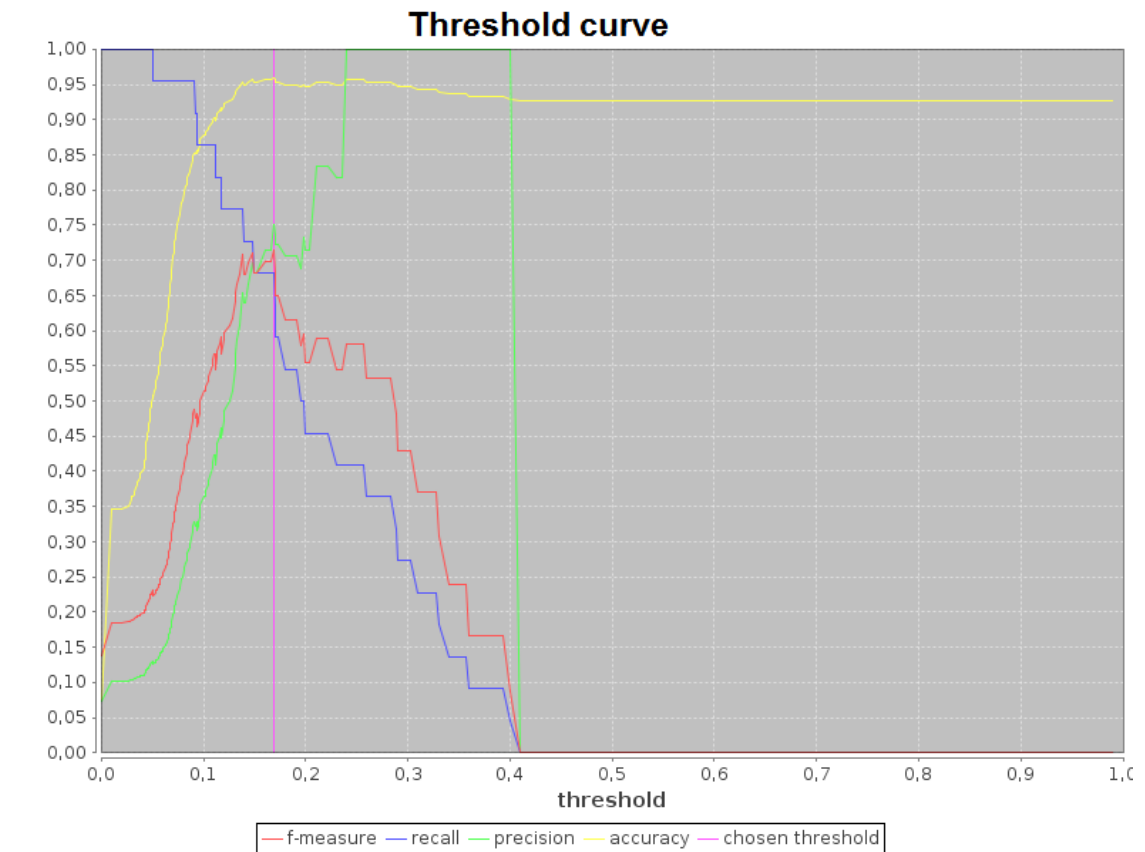


Abbildung 3.5: Beispiel für vom Evaluator ausgegebene Kurven: Schwellenwertkurve (oben) und ROC-Kurve (unten), wobei $FPR = \text{False Positive Rate} = \frac{FP}{FP+TN}$ und $TPR = \text{True Positive Rate} = \frac{TP}{TP+FN}$

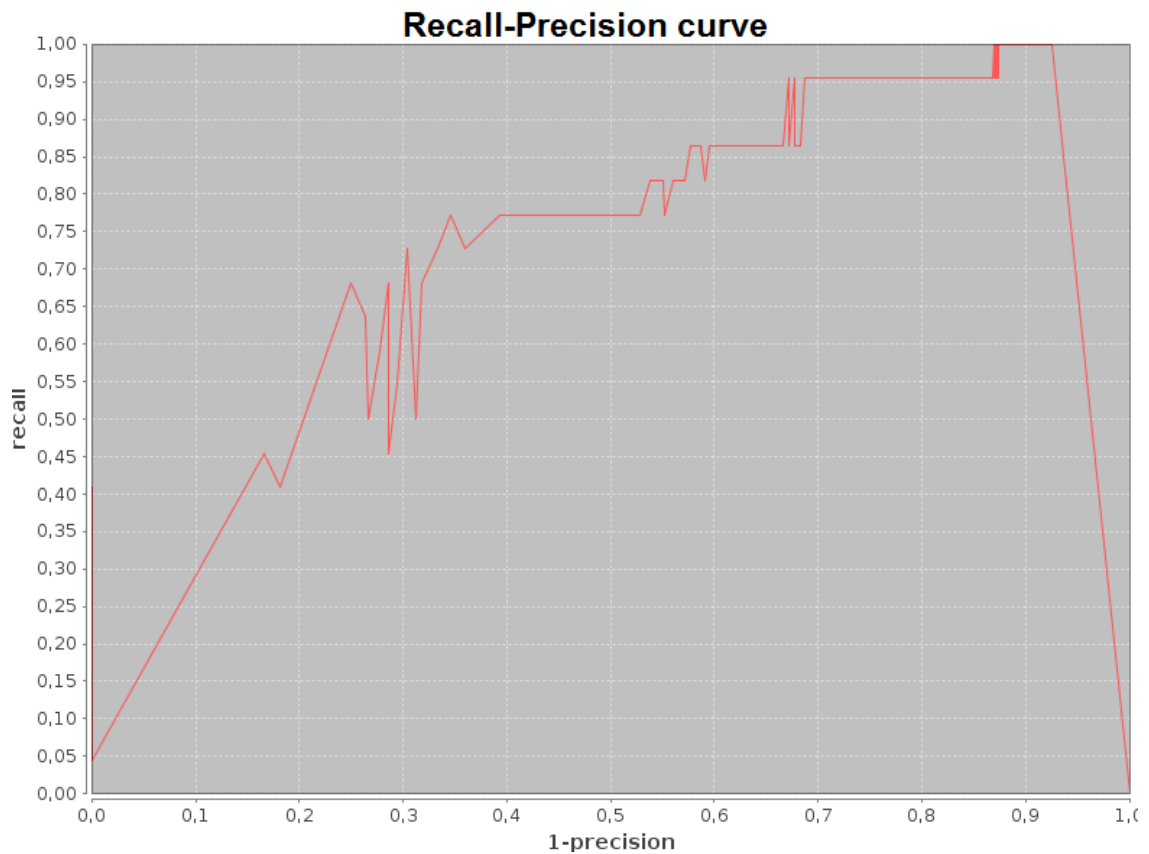


Abbildung 3.6: Beispiel für vom Evaluator ausgegebene Recall-Precision Kurve

```
# This file was constructed on Sat Sep 04 11:44:46 CEST 2010 by class pipeline2.annotator.CrossValidator with the
# following properties:
# WordNet version: 3.0 cat: 03
# Wikipedia version: 2009.08.22 cat: 03
# PPR Annotator graph version: 30
# PPV similarity measure: chi2
# Filter activated: false
# Bonus activated: false
# Maximum number of articles assigned to one synset: 1
#####
# Cross-Validation:
# Fold No.: Threshold ; Recall ; Precision ; Accuracy ; F-Measure ; TP ; TN ; FP ; FN
# Fold 0: 0,2250 ; 0,7391 ; 0,5484 ; 0,9029 ; 0,6296 ; 17 ; 169 ; 14 ; 6
# Fold 1: 0,2663 ; 0,9091 ; 0,6452 ; 0,9274 ; 0,7547 ; 20 ; 146 ; 11 ; 2
# Fold 2: 0,2250 ; 0,9583 ; 0,7188 ; 0,9213 ; 0,8214 ; 23 ; 94 ; 9 ; 1
# Fold 3: 0,2250 ; 0,8636 ; 0,6129 ; 0,9250 ; 0,7170 ; 19 ; 166 ; 12 ; 3
# Fold 4: 0,2890 ; 0,8750 ; 0,6774 ; 0,9293 ; 0,7636 ; 21 ; 150 ; 10 ; 3
# Fold 5: 0,2250 ; 0,8500 ; 0,5484 ; 0,9050 ; 0,6667 ; 17 ; 145 ; 14 ; 3
# Fold 6: 0,2250 ; 0,8571 ; 0,5806 ; 0,8904 ; 0,6923 ; 18 ; 112 ; 13 ; 3
# Fold 7: 0,2274 ; 0,7692 ; 0,6452 ; 0,8944 ; 0,7018 ; 20 ; 124 ; 11 ; 6
# Fold 8: 0,2148 ; 0,7826 ; 0,5625 ; 0,8876 ; 0,6545 ; 18 ; 132 ; 14 ; 5
# Fold 9: 0,2250 ; 0,9091 ; 0,6452 ; 0,9508 ; 0,7547 ; 20 ; 231 ; 11 ; 2
# Average: 0,2348 ; 0,8513 ; 0,6184 ; 0,9134 ; 0,7156
# Average2: 0,8502 ; 0,6186 ; 0,9157 ; 0,7161
# Std. Deviation: 0,0235 ; 0,0691 ; 0,0578 ; 0,0205 ; 0,0585
```

Abbildung 3.7: Beispiel für ein Ausgabefile des CrossValidators

4 Evaluation

Für die Evaluation der vorgestellten Methoden steht ein Gold-Standard mit 1815 Paaren zur Verfügung. Davon sind 227 als positiv und 1588 als negativ zu klassifizieren. Aufgrund dieser ungleichen Verteilung macht die Accuracy, welche den Anteil der korrekten unter sämtlichen Klassifizierungen angibt, als Evaluationsmaß wenig Sinn. Schon bei einer Klassifikation aller 1815 Paare als negativ würde eine Accuracy von 0.875 erreicht. Die Accuracy berechnet sich nach folgender Formel mit den Abkürzungen TP = True Positives, TN = True Negatives, FP = False Positives und FN = False Negatives:

$$Acc = \frac{TP+TN}{TP+FP+TN+FN}$$

Aus diesem Grund haben wir uns für das F-Measure als Evaluationsmaß entschieden. Dieses stellt das gewichtete harmonische Mittel von Genauigkeit (Precision) und Trefferquote (Recall) dar. Im Unterschied zur Accuracy ist die Klassifikation sämtlicher Beispiele als negativ hier nicht zielführend und ergibt den Wert 0. Das F-Measure berechnet sich nach folgender Formel:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$
$$(precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN})$$

Bei der Zusammensetzung des Gold-Standards wurde darauf geachtet eine möglichst ausgewogene Datenmenge zu schaffen, die verschiedene Gebiete abdeckt und dadurch repräsentativ wird. Die manuellen Zuweisungen des Gold-Standards wurden von drei menschlichen Annotatoren mit einer durchschnittlichen Übereinstimmung von 0.972 vorgenommen. Die Evaluierung der vorgestellten Ansätze haben wir zunächst auf den Trainingsdaten selbst und anschließend mit einer 10-fold Cross-Validation mit verschiedenen Einstellungen durchgeführt. Ziel der Evaluation ist es, die neuwertige semantische PageRank Methode (Abschnitt 2.2) mit der rein String-basierten Methode (Abschnitt 2.1) zu vergleichen und den Nutzen einer kombinierten Methode (Abschnitt 2.4) abzuschätzen. Als zusätzliche Baseline wird eine kontextunabhängige zufalls-basierte Methode herangezogen, die eine Klassifikation auf Basis zufälliger Ähnlichkeitswerte durchführt. Außerdem sollen die verschiedenen Ähnlichkeitsmaße (Abschnitt 2.3), die Auswirkungen verschiedener Zusammensetzungen der Bags of Words und die Stabilität der Verfahren bei der Cross-Validation untersucht werden.

Die wohl offensichtlichste Methode, die Ergebnisse zu verbessern, ist es, die Zusammensetzung der Bags of Words zu optimieren. Wir sind bereits in Abschnitt 1.3 darauf eingegangen, dass die Einbeziehung verwandter Synsets/Artikel sich positiv auf die Beschreibung der jeweiligen Bedeutung auswirken kann. Insbesondere bei WordNet Synsets enthalten die Bags of Words häufig nur sehr wenige Wörter. Es macht daher Sinn den aus Synonymen, Gloss und Beispiel bestehenden Synset Text um stark verwandte Synsets, also Hyponyme und/oder Hyperonyme, zu erweitern, um so die Bedeutung besser zu umschreiben. Ebenso kann es sinnvoll sein, die Bags of Words von Wikipedia Artikeln nicht ausschließlich aus dem ersten Absatz und Titel aufzubauen, sondern Redirects und/oder Kategorien des entsprechenden Artikels mit aufzunehmen. Die Evaluation zeigt, dass durch entsprechende Variationen der Bags of Words deutliche Verbesserungen der Performanz erzielt werden können. Folgende Variationen wurden bei der Evaluation getestet:

WordNet:

- S: Synset (Synonyme, Gloss, Beispiel)
- SD: Direkte Nutzung des Synsets ohne Bag of Words
- HE: Hyperonym Synsets (Synonyme, Gloss, Beispiel)
- HO: Hyponym Synsets (Synonyme, Gloss, Beispiel)

Wikipedia:

- T: Artikel Titel
- P: Erster Absatz des Artikels
- P2: Erste zwei Absätze des Artikels
- R: Redirects
- C: Categories

Des Weiteren ist in den Trainingsdaten zu beobachten, dass Paare deren Wikipedia Artikel Titel exakt mit einem der WordNet Synset Synonyme übereinstimmt, mit einer hohen Wahrscheinlichkeit eine Bedeutung teilen. So stimmen z.B. das Synset „S: (n) soul (the human embodiment of something) the soul of honor“ und der Wikipedia Artikel mit dem Titel „Soul“ überein. Nicht die gleiche Bedeutung haben jedoch das gleiche Synset und der Wikipedia Artikel „Soul music“, was der Heuristik entspricht, da der Titel nicht vollständig mit dem Synonym „soul“ übereinstimmt. Die Heuristik kann jedoch in einigen Fällen auch Fehler verursachen: Das Synset „S: (n) ditch (a long narrow excavation in the earth)“ stimmt beispielsweise mit dem Artikel „Ditch (fortification)“ überein, nicht jedoch mit dem Artikel „Ditch“. Dennoch ist auf dem Gold-Standard mit dieser Heuristik alleine bereits ein F-Measure von 0.695 zu erreichen.

Entsprechend dieser Beobachtung haben wir ein Bonussystem eingeführt, das die berechneten Ähnlichkeitswerte von entsprechenden Paaren mit einem Faktor multipliziert. Die Evaluation zeigt, dass dadurch eine deutliche Verbesserung der Performanz erreicht werden kann. Die Ergebnisse mit Bonussystem sollten jedoch mit Vorsicht betrachtet werden, denn ein Synonym ist in der Regel in mehreren Synsets enthalten. Dadurch wird ein Artikel dessen Titel mit einem Synonym übereinstimmt möglicherweise mehreren Synsets zugewiesen, was in der Regel nicht korrekt ist. Das Synonym „soul“ kommt zum Beispiel in gleich fünf verschiedenen Synsets vor. Das Bonussystem würde folglich bei einem Vergleich mit dem Artikel „Soul“ den Bonus an alle fünf Synsets vergeben, obwohl hier natürlich nur ein Synset korrekt ist. Es ist also möglich, dass die guten Ergebnisse auf den Trainingsdaten mit einer zu starken Anpassung an diese zusammenhängen.

Um einen optimalen Bonus zu finden, wird (parallel zum Training des Schwellenwertes) ein Training für den Bonus durchgeführt. Dabei ergeben sich (insbesondere beim String-basierten Ansatz) häufig relativ hohe Faktoren (zwischen 3 und 4), was zur Folge hat, dass die Zuweisung prinzipiell von dem Bonussystem vorgenommen wird, da bei solch hohen Faktoren die Vergabe eines Bonus in der Mehrheit der Fälle direkt zu einer positiven Klassifikation des entsprechenden Beispiels führt. Die eigentlich betrachteten Methoden (PageRank-basiert, String-basiert und Kombination) haben in einem solchen extremen Fall dann nur noch Einfluss, wenn die Anzahl der einem Synset zugewiesenen Artikel begrenzt wird (siehe nächster Absatz) und mehrere Paare mit dem gleichen Synset einen Bonus erhalten: In diesem Fall entscheidet eine der drei eigentlichen Methoden welches der Beispiele positiv klassifiziert wird (sofern die maximale Anzahl auf eins beschränkt ist).

Im Gold-Standard enthalten sind 320 WordNet Synsets mit durchschnittlich 5,67 möglichen zugehörigen Wikipedia Artikeln. Davon entsprechen im Schnitt jedoch nur 0,71 Artikel einer korrekten Zuordnung. Dabei gibt es 99 Synsets ohne zugehörigen Artikel, 215 Synsets mit genau einem Artikel und 6 Synsets mit genau zwei zugehörigen Artikeln. Die Tatsache, dass 98,125% der Synsets aus dem Gold-Standard maximal einen zugehörigen Artikel besitzen, kann dazu genutzt werden, die Ergebnisse zu verbessern. So kann, sofern für ein Synset mehrere mögliche Artikel den Schwellenwert übertreffen, nur der Artikel mit dem höchsten Ähnlichkeitswert zugeordnet werden. Die Implementation ermöglicht es, die maximale Anzahl an zugewiesenen Artikeln entsprechend zu begrenzen. Abbildung 4.1 zeigt beispielhaft die Entwicklung der F-Measure Werte des PageRank-basierten Verfahrens (Ähnlichkeitsmaß χ^2 , Bag of Words Zusammensetzung S+HE bzw. T+P+R+C) bei einer Variation der maximalen Anzahl einem Synset zugewiesener Artikel. Zusätzlich zu den Werten mit und ohne Bonus wird die Entwicklung des Schwellenwertes (ohne Bonussystem) angezeigt. Es ist zu sehen, dass die Performanz mit einer Steigerung der maximalen Anzahl an Zuweisungen pro Synset abnimmt. Der gewählte Schwellenwert steigt unterdessen, da in diesem Fall weniger Beispiele durch die Begrenzung zurückgewiesen werden und somit der Schwellenwert diese Aussortierung vornehmen muss. Bemerkenswert ist zudem, dass bei Verwendung des Bonussystems die Performanz weniger stark abnimmt als ohne Bonus, was damit zu erklären ist, dass in der Regel nur wenige Paare eines Synsets einen Bonus erhalten. Den Auswertungsdaten ist zu entnehmen, dass gleichzeitig der gewählte Bonusfaktor zunimmt, sodass dem Bonussystem mehr Beachtung zukommt. Um die Ergebnisse zu optimieren, haben wir in unseren folgenden Experimenten die Anzahl an Zuweisungen auf eins beschränkt.

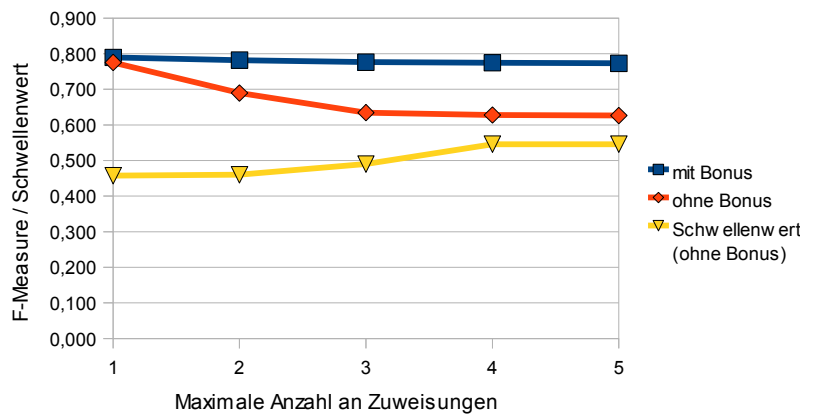


Abbildung 4.1: Beschränkung der Anzahl einem Synset zugewiesener Artikel

4.1 Analyse und Bewertung der Ergebnisse

Bevor wir ausführlich auf die Performanz der verschiedenen Methoden und Variationen eingehen, wollen wir einen kurzen Blick auf die zeitliche Performanz der Algorithmen werfen. Der Arbeitsaufwand der Methoden geht aus den beiden in Kapitel 3 eingeführten Pipelines hervor. Das grundsätzlich notwendige Preprocessing (Erstellung der Bags of Words) benötigt nur wenige Sekunden pro Synset bzw. Artikel, hängt aber natürlich davon ab, ob noch verwandte Synsets bzw. Redirects und/oder Kategorien einbezogen werden und kann bei größeren Datensätzen wie dem hier verwendeten Gold-Standard auch über eine Stunde dauern (bei einem 3GHz Prozessor). Den größten Aufwand stellt jedoch die Berechnung der PageRank Vektoren dar, welche allerdings für die String-basierte Methode entfällt. Für die Berechnung eines einzelnen PageRank Vektors können bei einem 3 GHz Prozessor grob 15 Sekunden veranschlagt werden, was sich bei größeren Datensätzen schnell zu mehreren Stunden aufsummieren kann. Der Aufwand für die Berechnung eines Ähnlichkeitswerts ist verglichen mit den vorherigen Schritten fast vernachlässigbar. Für den String-basierten Ansatz ist dieser Schritt, selbst bei größeren Datensätzen, nach einigen Sekunden abgeschlossen. Die PageRank-basierte Methode benötigt etwas länger, da die PageRank Vektoren mit jeweils 117.659 Einträgen relativ groß sind. Es ist also festzustellen, dass der Arbeitsaufwand für die Zuweisung von WordNet Synsets und Wikipedia Artikeln mit der PageRank-basierten Methode erheblich größer ist als mit der String-basierten Methode. Andererseits ist es jedoch

möglich die PageRank Vektoren sämtlicher WordNet Synsets und Wikipedia Artikel vorzuberechnen. Anschließend können sie dank ihrer einheitlichen Repräsentation für die Berechnung von Ähnlichkeiten in jeder beliebigen Paarung genutzt werden, sodass die Berechnung von weiteren PageRank Vektoren wegfällt.

Bei der Analyse und Bewertung der Ergebnisse beziehen wir uns im Folgenden auf die Tabellen 4.1 und 4.2. Tabelle 4.1 zeigt die Ergebnisse (F-Measure), trainiert und ausgewertet auf dem Gold-Standard (Abkürzung: T), was der maximal möglichen Performanz auf den Trainingsdaten entspricht, und mit 10-fold Cross-Validation (Abkürzung: CV) mit den verschiedenen in Kapitel 2 vorgestellten Ansätzen und Ähnlichkeitsmaßen. Für die Werte in dieser Tabelle wurde das oben beschriebene Bonussystem nicht genutzt. Es wurden unterschiedliche Zusammensetzungen der Bags of Words getestet, die Abkürzungen dazu sind der Übersicht in der Einleitung dieses Kapitels zu entnehmen. Die maximale Anzahl an einem Synset zugewiesenen Wikipedia Artikeln wurde auf eins beschränkt. Tabelle 4.2 beschreibt entsprechend die Ergebnisse unter Anwendung des Bonussystems. In den Tabellen sind für jeden der drei Ansätze, jede Variation der WordNet Bags of Words und für die Auswertung auf den Trainingsdaten bzw. mit Cross-Validation jeweils die besten Ergebnisse hervorgehoben.

Die zufalls-basierte Methode, deren Werte unabhängig von der Zusammensetzung der Bags of Words sind, erreicht auf den Trainingsdaten ein F-Measure von etwa 0.53 (ohne Bonus) bzw. 0.63 (mit Bonus). Diese Werte werden von allen nicht zufalls-basierten Methoden übertroffen, sodass sich für jede dieser Methoden ein gewisser Nutzen ergibt. Wir wollen uns zunächst mit einer Analyse der Ergebnisse ohne Bonussystem befassen, bevor wir eine Bewertung der Ergebnisse mit Bonussystem vornehmen.

4.1.1 Ergebnisse ohne Bonussystem

Die folgenden Aussagen beziehen sich zunächst auf die auf den Trainingsdaten ausgeführte Evaluation ohne Bonussystem (siehe Tabelle 4.1). Mit teilweise über 10% Differenz zwischen den einzelnen Ähnlichkeitsmaßen unterscheiden sich die Ergebnisse des auf dem Personalisierten PageRank Algorithmus basierenden Ansatzes relativ stark. Die Rangfolge der einzelnen Maße ist für verschiedene Bags of Words Zusammensetzungen jedoch weitgehend stabil. Die besten Werte erreichen wir für nahezu alle Bag of Words Kombinationen mit dem gewichtenden Maß χ^2 . Für aus Synset + Hyperonym bzw. Artikel Titel + erster Absatz + Redirects + Kategorien bestehende Bags of Words ist mit diesem Maß ein F-Measure von bis zu 0.776 möglich. Über alle Variationen betrachtet deutlich schlechter schneiden die Maße Intersection (maximal 0.745) und Cosinus Distanz (maximal 0.753) ab. Das Schlusslicht bilden schließlich Euklidische Distanz (maximal 0.686) und Punktprodukt (maximal 0.694). Sowohl bei der Euklidischen Distanz als auch beim Punktprodukt wirken sich die fehlenden Gewichtungen, wie sie bei χ^2 bzw. Cosinus Distanz existieren, negativ auf die Ergebnisse aus. Die Unterschiede zwischen den Ähnlichkeitsmaßen wurden bereits in Abschnitt 2.3 diskutiert.

Für die Zusammensetzung der Bags of Words kristallisiert sich heraus, dass für WordNet Synsets insbesondere die Hinzunahme von Hyperonymen von Vorteil ist. Dies hängt zunächst damit zusammen, dass die in dem Gold-Standard verwendeten Synsets häufig auf unterster Ebene liegen und somit gar keine Hyponyme aufweisen (z.B. die Synsets zu den Wörtern „seedling“ und „fife“). Ein Blick in den verwendeten Gold-Standard offenbart, dass von den 320 Synsets ganze 243 kein Hyponym besitzen, was einem Anteil von 0.759 entspricht. Dies erklärt den vergleichsweise geringen Einfluss bei Einbeziehung der Hyponyme auf die Ergebnisse. Weitere Probleme können dadurch entstehen, dass ein Synset (sofern es nicht auf unterster Ebene liegt) oft mehrere Hyponyme hat, welche die Bags of Words möglicherweise zu stark dominieren. Im Unterschied dazu hat jedes Synset genau ein Hyperonym. Nehmen wir beispielsweise das Synset „S: (n) actor, histrion, player, thespian, role player (a theatrical performer)“.

Das Synset hat wie immer genau ein Hyperonym und ganze 91 Hyponyme, was damit zusammenhängt, dass hier eine große Menge an Schauspielern vorkommt, die zur Umschreibung des Synsets „actor“ allerdings kaum einen positiven Beitrag leisten und zudem die Wörter des eigentlich betrachteten Synsets so stark dominieren, dass diese kaum noch Einfluss haben. Auch wenn dies ein sehr extremes Beispiel ist, wird deutlich, dass das Hyperonym als Oberbegriff häufig eine passendere Umschreibung liefert als die Hyponyme. Nimmt man das Synset „S: (n) Depardieu, Gerard Depardieu (French film actor (born in 1948))“, so eignet sich das Hyperonym „actor“ durchaus gut als zusätzliche Umschreibung. Das Beispiel „actor“ ist allerdings wohl eher als Ausnahmefall anzusehen. In der Regel ist auch die Einbeziehung von Hyponymen für die Performanz von Vorteil, was die Auswertungsdaten auch bestätigen, wenn auch in nur geringem Maße, da nur ca. $\frac{1}{4}$ der Synsets im Gold-Standard überhaupt ein Hyponym besitzt.

Bei Wikipedia Artikeln wirkt sich insbesondere die Aufnahme von Kategorien positiv auf die Ergebnisse aus. Geringeren Nutzen bringen Redirects, die sicherlich hilfreich sind, jedoch zu wenig ins Gewicht fallen, da die Bags of Words von Wikipedia Artikeln in der Regel deutlich mehr Wörter enthalten als beispielsweise bei WordNet. Interessant ist in der Hinsicht auch, dass die Erweiterung des betrachteten Artikel Texts von einem auf zwei Absätze die Ergebnisse deutlich verschlechtert, was vermutlich damit zusammenhängt, dass der zweite Absatz den betrachteten Artikel häufig schon zu allgemein beschreibt und dadurch den Einfluss der wirklich wichtigen Wörter schmälert.

Die direkte Nutzung der WordNet Synsets (Abkürzung: SD) ohne Berechnung einer Bag of Words bringt hinsichtlich der Performanz nur dann einen Vorteil, wenn keine Hyponyme bzw. Hyperonyme in die Bags of Words einbezogen werden. Unter Hinzunahme der Wikipedia Kategorien ist damit ein F-Measure von bis zu 0.754 möglich, während die gewöhnliche Methode (Erstellung von Bags of Words für WordNet Synsets, Abkürzung: S) nur Werte bis zu 0.726 erreicht. Ein Vergleich zu den WordNet Bags of Words, welche Hyperonyme bzw. Hyponyme einbeziehen, ergibt keine Vorteile für die direkte Gewichtung des jeweils betrachteten Synsets. Es ist jedoch möglich, dass eine Methode, die neben dem betrachteten Synset auch benachbarte Synsets direkt gewichtet, eine bessere Performanz erreichen kann. Die durch den Wegfall von zu erstellenden Bags of Words für WordNet Synsets entstehenden Effizienzvorteile sind vor dem Hintergrund der deutlich aufwändigeren Berechnung der PageRank Vektoren allerdings vernachlässigbar.

Bei einer Auswertung mit 10-fold Cross-Validation ergeben sich erwartungsgemäß jeweils leicht schlechtere Ergebnisse, wobei die zuvor getroffenen Feststellungen (bezüglich Ähnlichkeitsmaßen und Zusammensetzung der Bags of Words) bestehen bleiben. Das höchste erreichbare F-Measure beträgt hier 0.761 (statt 0.776), im Durchschnitt ergibt sich für das beste Ähnlichkeitsmaß χ^2 eine Verschlechterung der Ergebnisse um 0.010, was auf eine relativ hohe Stabilität eines einmal trainierten Schwellenwertes auf von den Trainingsdaten verschiedenen Testdaten schließen lässt. Somit ist davon auszugehen, dass das Verfahren auch für die großflächige Synset-Artikel Zuweisung geeignet ist.

Erwartungsgemäß übertrifft der in dieser Arbeit vorgestellte semantische, auf dem Personalisierten PageRank Algorithmus basierende Ansatz den als Baseline dienenden String-basierten Ansatz mit dem als am besten geeigneten Ähnlichkeitsmaß χ^2 deutlich. Man könnte nun einwenden, dass für den String-basierten Ansatz lediglich mit dem Ähnlichkeitsmaß Cosinus Distanz berechnete Werte vorliegen. Vergleicht man den String-basierten Ansatz folglich mit den auf der Cosinus Distanz basierenden Werten des Personalisierten PageRank Ansatzes, lässt sich tatsächlich kein eindeutiger Sieger ausmachen: Je nach Zusammensetzung der Bags of Words ist mal die String-basierte Methode mal der Personalisierte PageRank Algorithmus überlegen, wobei der PageRank Ansatz zumindest den deutlich besseren Maximalwert erreicht (auf den Trainingsdaten 0.753 gegenüber 0.738). Man könnte nun also versuchen zu argumentieren, dass die Werte der String-basierten Methode mit dem Ähnlichkeitsmaß χ^2 ebenfalls besser sein werden. Dies ist jedoch ein Trugschluss, da ein Ähnlichkeitsmaß nicht allgemein, sondern immer nur in

Abhängigkeit der Methode, von der es genutzt wird, bewertet werden sollte. Wie bereits in Abschnitt 2.3 erläutert, eignet sich das gewichtende Maß χ^2 nicht für die String-basierte Methode, deren Vektoren ausschließlich aus den Werten 0 und 1 bestehen.

Die Feststellung, dass sich insbesondere die Hinzunahme von Hyperonymen und Wikipedia Kategorien in die Bags of Words positiv auf die Ergebnisse auswirkt, hat auch für die String-basierte Methode Bestand. Bei Nutzung der Cross-Validation verschlechtern sich die Ergebnisse durchschnittlich um 0.008, was vergleichbar mit dem für den PageRank ermittelten Wert 0.010 ist. Zusammenfassend ist also festzustellen, dass die semantische auf dem Personalisierten PageRank basierende Methode bei ähnlicher Stabilität zu deutlich verbesserten Ergebnissen führt. Die Maximalwerte (Trainingsdaten/Cross-Validation) liegen bei 0.776/0.761 (PageRank) bzw. 0.738/0.735 (String-basiert).

Als dritten Ansatz betrachten wir die Kombination der String-basierten Methode mit dem Personalisierten PageRank, wie sie in Abschnitt 2.4 vorgestellt wurde. Dabei wurde für das PageRank-basierte Verfahren jeweils das performanteste Ähnlichkeitsmaß verwendet (bis auf wenige Ausnahmen also χ^2). Über die verschiedenen Zusammensetzungen der Bags of Words betrachtet findet die Kombination ihre Daseinsberechtigung. Auf den Trainingsdaten werden für nur 4 von 20 Bags of Words Zusammensetzungen von der einfachen PageRank Methode bessere Werte erzielt. Der Maximalwert liegt auf den Trainingsdaten mit 0.781 nochmals über dem für PageRank gemessenen Wert von 0.776. Ein genauerer Blick in die Auswertungsdaten offenbart dabei, dass sich durch Anwendung der Kombination wie erwartet insbesondere der Precision Wert steigern lässt (auf Kosten des Recalls). Dies hängt damit zusammen, dass ein Beispiel nur unter erschwerten Bedingungen positiv klassifiziert wird, wodurch sich die Anzahl an False Positives verringert, während die Anzahl an False Negatives zunimmt.

Bei den einzelnen Methoden erreichen die String-basierte Methode und die PageRank-basierte Methode auch bei ähnlichen F-Measure Werten verschiedene Recall und Precision Werte. Daran ist bereits zu erkennen, dass hier unterschiedliche Fehler gemacht werden, was die Kombinationsmethode ausnutzt. Diese Fehler hängen überwiegend damit zusammen, dass der PageRank-basierte Ansatz semantisch ist und der String-basierte Ansatz nicht. Man kann dazu sehr gut das Beispiel aus Abschnitt 2.1 heranziehen. So wird das Synset „Johannesburg“ und der korrespondierende Artikel „Johannesburg“ von der String-basierten Methode (ohne Einbeziehung weiterer Wörter und ohne Bonussystem) fälschlicherweise als negativ klassifiziert, weil nur drei Wörter übereinstimmen. Der PageRank-basierte Ansatz klassifiziert das Beispiel dagegen korrekt, da dieser den semantischen Zusammenhang der Wörter „large economy“ und „commercial center“ erkennt. Für den umgekehrten Fall betrachten wird das folgende Synset-Artikel Paar:

- Wikipedia: Judas is a manga by Suu Minazuki. There are a total of five volumes in this series. The first was published in English by Tokyopop on October 10, 2006. The second volume of this series was released by Tokyopop on February 13, 2007. Judas is cursed for his sins to kill six hundred and sixty six (666) people to regain his humanity. However, he is forbidden human contact and has no corporeal body. In order to kill, he uses his slave, Eve, to kill for him. Every time Eve's blood is spilled, Judas comes out and forces Eve to „say his prayers“, in other words, kill...
- WordNet: Jude, Saint Jude, St. Jude, Judas, Thaddaeus ((New Testament) supposed brother of St. James; one of the Apostles who is invoked in prayer when a situation seems hopeless)

Die String-basierte Methode klassifiziert das Beispiel korrekt negativ, da lediglich ein Wort („Judas“) übereinstimmt. Fälschlicherweise positiv bewertet wird das Beispiel hingegen von der PageRank-basierten Methode, die scheinbar einen semantischen Zusammenhang zwischen den Texten sieht. Dieser kann beispielsweise aus den vielen Zahlenangaben („five“, „six“, „one“) resultieren, welche alle das gleiche Hyperonym besitzen.

Bei der Cross-Validation ist die einfache PageRank Methode in nur 6 von 20 Bag of Words Variationen besser als die Kombinationsmethode, der Maximalwert liegt mit 0.762 allerdings nur unwesentlich über dem Maximalwert von PageRank alleine (0.761). Der Nutzen der Kombinationsmethode ist somit nicht als besonders hoch einzuschätzen, was auch damit zusammenhängt, dass hier eine sehr einfache Methode zur Kombination genutzt wurde. In der Fehleranalyse (Abschnitt 4.2) werden mögliche Verbesserungen diskutiert.

4.1.2 Ergebnisse mit Bonussystem

Die Ergebnisse unter Nutzung des Bonussystems sind Tabelle 4.2 zu entnehmen. Es ist zu beobachten, dass sich die Werte zwischen den einzelnen Ähnlichkeitsmaßen für den PageRank-basierten Ansatz weniger stark unterscheiden als ohne Bonus. Insbesondere die Werte von χ^2 , Intersection und Cosinus Distanz liegen sehr dicht beieinander, was darauf schließen lässt, dass der Bonus bei der Berechnung eine dominierende Rolle einnimmt. So entscheiden die eigentlichen Verfahren nur dann über eine positive bzw. negative Klassifikation, wenn mehrere Paare eines Synsets einen Bonus erhalten (aufgrund der Beschränkung der Anzahl einem Synset zugewiesener Artikel auf eins). Daraufhin deutet auch, dass sich die Werte für die verschiedenen Bags of Words Zusammensetzungen weniger stark unterscheiden. Der Maximalwert für das PageRank-basierte Verfahren mit Bonus beträgt 0.790 mit dem Ähnlichkeitsmaß χ^2 (auf den Trainingsdaten). Dazu werden in den Bags of Words Hyperonyme bzw. Redirects und Kategorien hinzugefügt. Wie auch schon bei der Auswertung ohne Bonussystem festgestellt, bringt die direkte Nutzung der WordNet Synsets ohne Bag of Words Berechnung nur einen Vorteil gegenüber den WordNet Bags of Words ohne Hinzunahme verwandter Synsets (Abkürzung: S). Die Auswertung mit Cross-Validation offenbart mit einem durchschnittlichen Rückgang der Werte (über die verschiedenen Bag of Words Zusammensetzungen) um 0.025 (für χ^2) eine stärkere Verschlechterung der Ergebnisse als ohne Bonussystem (Rückgang: 0.010). Dies lässt auf eine geringere Stabilität des Bonussystems auf großen Datensätzen schließen.

Interessant im Vergleich zu den Ergebnissen ohne Bonussystem ist, dass bei der Nutzung eines Bonus die PageRank-basierte Methode der String-basierten Methode nicht mehr überlegen ist. In vielen Fällen ergeben sich mit der String-basierten Methode hier sogar bessere Werte. So liegt der Maximalwert mit 0.797 auch leicht über dem Maximalwert des PageRank-basierten Ansatzes (0.790). Es ist dabei zu beobachten, dass die Bonusfaktoren bei der String-basierten Methode mit Werten zwischen 3 und 4 besonders hoch liegen. Die Ursache ist also vor allem in der dominierenden Rolle des Bonussystems zu suchen. So verliert auch die Kombination mit Bonussystem an Wert, da diese häufig schon von der String-basierten oder der PageRank-basierten Methode alleine übertroffen wird. Der Maximalwert der Kombination übersteigt mit 0.799 dennoch alle bisherigen Ergebnisse. Er wird erneut unter Hinzunahme von Hyperonymen und Kategorien erreicht.

Abschließend steht noch ein Vergleich zwischen den mit und ohne Bonussystem gemessenen Werten aus. Betrachtet man die Tabellen insgesamt, so fällt schnell ins Auge, dass das Bonussystem quer durch alle Ansätze und Bags of Words Kombination auf den Trainingsdaten zu erheblichen Verbesserungen von ca. 5% führt. Dabei darf jedoch nicht die in der Einleitung des Kapitels vorgetragene Kritik am Bonussystem vergessen werden.

4.2 Fehleranalyse

Es ist zunächst festzustellen, dass die betrachteten Methoden mit F-Measure Werten von knapp unter 80% bereits sehr gute Ergebnisse erzielen. Lässt man die Begrenzung der Anzahl an Zuweisungen pro Synset und das Bonussystem weg, so offenbart sich dennoch ein gewisser Verbesserungsbedarf. Fehler-

WordNet	Wikimedia	Personalisierter PageRank																							
		String-basiert		cos						int						χ^2				euc		dot		Kombination	
		T	CV	T	CV	T	CV	T	CV	T	CV	T	CV	T	CV	T	CV	T	CV	T	CV	T	CV		
S	T+P	.691	.675	.690	.687	.706	.694	.708	.699	.653	.650	.649	.631	.727	.703										
S	T+2P	.677	.674	.658	.657	.690	.687	.697	.686	.673	.668	.629	.608	.725	.715										
S	T+P+R	.684	.683	.686	.685	.709	.703	.707	.704	.665	.665	.631	.625	.725	.722										
S	T+P+C	.698	.698	.690	.686	.708	.701	.726	.684	.680	.680	.634	.608	.743	.741										
S	T+P+R+C	.699	.698	.692	.689	.693	.683	.719	.715	.680	.680	.636	.614	.734	.730										
SD	T+P	-	-	.649	.645	.696	.690	.719	.709	.627	.605	.624	.618	-	-										
SD	T+2P	-	-	.646	.637	.676	.665	.686	.665	.647	.624	.620	.614	-	-										
SD	T+P+R	-	-	.650	.640	.697	.681	.721	.720	.626	.607	.610	.607	-	-										
SD	T+P+C	-	-	.675	.672	.704	.693	.754	.739	.634	.613	.618	.610	-	-										
SD	T+P+R+C	-	-	.664	.658	.687	.672	.736	.715	.649	.638	.614	.606	-	-										
S+HE	T+P	.726	.711	.738	.704	.737	.726	.756	.738	.664	.664	.686	.648	.774	.737										
S+HE	T+2P	.719	.694	.725	.719	.720	.704	.732	.723	.675	.674	.672	.668	.720	.708										
S+HE	T+P+R	.719	.717	.731	.718	.740	.724	.762	.757	.661	.657	.690	.689	.755	.753										
S+HE	T+P+C	.738	.735	.752	.726	.745	.741	.765	.744	.683	.674	.694	.683	.781	.762										
S+HE	T+P+R+C	.718	.703	.753	.743	.737	.730	.776	.761	.686	.678	.687	.686	.772	.755										
S+HO	T+P	.694	.668	.691	.689	.707	.703	.700	.689	.654	.650	.656	.638	.723	.688										
S+HO	T+2P	.678	.672	.668	.663	.698	.689	.690	.681	.676	.669	.635	.621	.722	.707										
S+HO	T+P+R	.689	.688	.683	.681	.705	.697	.698	.696	.658	.658	.635	.624	.725	.716										
S+HO	T+P+C	.702	.702	.701	.697	.702	.700	.722	.716	.676	.672	.636	.623	.756	.733										
S+HO	T+P+R+C	.695	.693	.700	.696	.692	.686	.711	.707	.665	.665	.636	.624	.727	.721										
S+HO+HE	T+P	.725	.708	.727	.708	.730	.715	.741	.721	.664	.659	.685	.685	.756	.723										
S+HO+HE	T+2P	.729	.725	.711	.710	.723	.714	.728	.728	.675	.669	.675	.671	.719	.717										
S+HO+HE	T+P+R	.727	.725	.726	.709	.733	.711	.747	.731	.653	.645	.681	.659	.761	.747										
S+HO+HE	T+P+C	.732	.724	.742	.726	.742	.741	.746	.727	.679	.668	.689	.678	.769	.749										
S+HO+HE	T+P+R+C	.724	.718	.734	.714	.737	.729	.762	.752	.679	.669	.677	.665	.769	.758										

Tabelle 4.1: Ergebnisse der automatischen Zuweisung ohne Bonussystem auf Trainingsdaten (T) und mit Cross Validation (CV), maximal ein Artikel pro Synset (Werte in F-Measure)

WordNet	Wikimedia	String-basiert		Personalisierter PageRank												Kombination			
		T	CV	cos			int			χ^2			euc			dot		T	CV
				T	CV	T	CV	T	CV	T	CV	T	CV	T	CV	T	CV		
S	T+P	.778	.772	.742	.726	.761	.740	.756	.718	.714	.684	.724	.706	.770	.754				
S	T+2P	.772	.746	.749	.735	.743	.728	.750	.720	.722	.692	.729	.715	.756	.746				
S	T+P+R	.776	.764	.735	.702	.759	.716	.753	.724	.708	.683	.713	.701	.761	.733				
S	T+P+C	.761	.731	.731	.705	.768	.747	.763	.756	.710	.680	.719	.693	.782	.753				
S	T+P+R+C	.779	.768	.747	.710	.774	.767	.768	.742	.716	.689	.722	.710	.769	.758				
SD	T+P	-	-	.716	.690	.752	.741	.760	.731	.724	.699	.703	.683	-	-				
SD	T+2P	-	-	.727	.712	.746	.721	.765	.741	.737	.720	.691	.650	-	-				
SD	T+P+R	-	-	.717	.690	.763	.741	.766	.747	.728	.702	.684	.663	-	-				
SD	T+P+C	-	-	.727	.711	.776	.767	.782	.756	.733	.732	.707	.687	-	-				
SD	T+P+R+C	-	-	.724	.699	.776	.768	.781	.753	.738	.707	.704	.679	-	-				
S+HE	T+P	.797	.787	.770	.720	.777	.756	.779	.745	.732	.722	.747	.723	.792	.779				
S+HE	T+2P	.781	.746	.784	.766	.770	.757	.774	.766	.734	.730	.739	.707	.794	.753				
S+HE	T+P+R	.783	.754	.773	.764	.780	.758	.781	.759	.705	.681	.741	.717	.770	.749				
S+HE	T+P+C	.796	.782	.784	.746	.781	.762	.785	.757	.719	.692	.739	.711	.799	.777				
S+HE	T+P+R+C	.793	.763	.786	.774	.786	.765	.790	.747	.719	.687	.746	.728	.796	.762				
S+HO	T+P	.778	.765	.737	.717	.750	.728	.751	.727	.705	.679	.728	.711	.761	.744				
S+HO	T+2P	.774	.763	.748	.734	.746	.737	.752	.727	.725	.698	.732	.713	.753	.742				
S+HO	T+P+R	.778	.768	.731	.687	.748	.721	.749	.710	.700	.693	.711	.684	.752	.742				
S+HO	T+P+C	.766	.739	.741	.708	.757	.734	.758	.743	.706	.686	.725	.713	.740	.740				
S+HO	T+P+R+C	.779	.771	.745	.696	.763	.738	.765	.758	.704	.680	.724	.718	.764	.755				
S+HO+HE	T+P	.797	.792	.770	.725	.771	.746	.765	.724	.719	.701	.749	.731	.785	.764				
S+HO+HE	T+2P	.787	.769	.772	.753	.773	.766	.775	.773	.738	.721	.742	.701	.779	.765				
S+HO+HE	T+P+R	.779	.752	.763	.711	.780	.779	.776	.762	.696	.667	.743	.738	.776	.750				
S+HO+HE	T+P+C	.791	.778	.777	.723	.776	.748	.778	.742	.712	.674	.742	.714	.789	.760				
S+HO+HE	T+P+R+C	.792	.774	.781	.775	.781	.761	.786	.756	.715	.684	.749	.727	.784	.766				

Tabelle 4.2: Ergebnisse der automatischen Zuweisung mit Bonussystem auf Trainingsdaten (T) und mit Cross-Validation (CV), maximal ein Artikel pro Synset (Werte in F-Measure)

quellen und daraus resultierende Verbesserungen sind an mehreren Stellen möglich:

Wir haben festgestellt, dass die Zusammensetzung der Bags of Words die Ergebnisse maßgeblich beeinflusst. Eine weitere Optimierung der Bags of Words ist daher naheliegend. Wie schon in Abschnitt 1.2.2 festgestellt, enthalten insbesondere Wikipedia Artikel häufig Wörter, die für die Artikelstruktur relevant sind, aber weniger dafür, den Inhalt zu beschreiben (z.B. Verweise auf Unterkapitel). Verbesserungen könnten also aus einer noch besseren Auswahl relevanter Wörter resultieren. Dabei sind auch Methoden denkbar, die beispielsweise „wichtige“ Wörter innerhalb eines Kontextes identifizieren und bei der Berechnung der PageRank Vektoren durch ein höheres Initialgewicht stärker gewichten. Möglich wäre zudem, statt dem ersten Absatz eines Artikels, alle ausgehenden Verweise in die Bag of Words aufzunehmen wie von Ponzetto and Navigli [2010] vorgeschlagen (siehe Kapitel 5). Weiteres Verbesserungspotential bieten die Ähnlichkeitsmaße. So hat sich gezeigt, dass sich zumindest für die PageRank basierte Methode ein gewichtendes Maß gut eignet. Es ist daher möglich, dass sich beispielsweise durch eine etwas anders gewählte Gewichtung Vorteile ergeben.

Die Liste der Fehlerquellen wird vom Personalisierten PageRank Algorithmus selbst ergänzt: Das Initialgewicht wird gleichmäßig auf sämtliche mögliche Bedeutungen der in den Bags of Words enthaltenen Wörter verteilt. Demnach erhalten auch Bedeutungen an Gewicht, die in dem Kontext, in dem das jeweilige Wort vorkommt, nicht korrekt sind. Zwar ist zu erwarten, dass diese falschen Bedeutungen eher isoliert im Graphen vorliegen, wodurch sich ihr Gewicht in den Weiten des Graphen verliert, dennoch entsteht ein gewisses Rauschen in der Repräsentation der entsprechenden Bags of Words. Dem Problem kann Abhilfe geschaffen werden, indem der Berechnung der PageRank Vektoren eine Word Sense Disambiguation vorangestellt wird, mit dem Ziel, die korrekten Bedeutungen zu identifizieren, um dann das Initialgewicht nur auf diese zu verteilen. Das Problem dabei ist, dass auch die Word Sense Disambiguation nach aktuellem Stand der Forschung zu viele Fehler macht, als dass durch diese Erweiterung eine deutliche Verbesserung zu erwarten ist.

Auch für das Kombinationsverfahren sind Optimierungen denkbar. So führt die Kombination in der jetzigen Form in erster Linie zu einer Zunahme der Precision, während der Recall entsprechend abnimmt. Es wäre daher auch möglich, dass durch ein auf die Kombination optimiertes Training der Schwellenwerte bessere Ergebnisse erzielt werden können (momentan werden die Schwellenwerte für die einzelnen Verfahren und nicht für die Kombination trainiert). Es ist dabei zu erwarten, dass die Schwellenwerte für die Kombination etwas niedriger ausfallen, als für die einzelnen Verfahren, da in der Kombination für eine positive Klassifikation zwei Verfahren zustimmen müssen. Anstatt nur eine positive Klassifikation vorzunehmen, wenn beide Methoden das Beispiel positiv klassifiziert haben, könnte man auch ein Voting Verfahren einführen, bei dem jede der beiden Methoden, je nachdem wie „sicher“ die jeweilige Klassifikation ist, unterschiedlich stark dafür abstimmt. Je größer die Differenz zwischen Ähnlichkeitswert des jeweiligen Beispiels und Schwellenwert ist, desto sicherer ist die Klassifikation.

In den Auswertungsdaten sind auf den ersten Blick keine klaren Muster in den Fehlern erkennbar. Ein tieferer Blick in die Daten offenbart, wie bereits in Abschnitt 4.1.1 festgestellt, dass Fehler der String-basierten Methode häufig mit einer zu geringen Überlappung der Wörter zusammenhängen und der PageRank-basierte Ansatz teilweise an Wörtern wie Zahlen scheitert, die semantisch eng miteinander verwandt sind, aber für die Gesamtbedeutung der Texte kaum von Relevanz sind. Häufig ist (insbesondere bei der String-basierten Methode) auch problematisch, dass die zu vergleichenden Texte von unterschiedlicher Länge sind. Nehmen wir beispielsweise das folgende Synset-Artikel Paar:

-
- Wikipedia: A schooner (pronounced /'sku:n?r/) is a type of sailing vessel characterized by the use of fore-and-aft sails on two or more masts with the forward mast being no taller than the rear masts. Schooners were first used by the Dutch in the 16th or 17th century, and further developed in North America from the early 18th century.
 - WordNet: schooner (sailing vessel used in former times)

Obwohl fast alle Wörter des WordNet Synsets im ersten Absatz des Wikipedia Artikels vorkommen, wird dieses Beispiel von der String-basierten Methoden fälschlicherweise negativ klassifiziert. Dies hängt damit zusammen, dass der erste Absatz des Artikels relativ lang ist und somit relativ gesehen nur wenige Wörter aus dem Artikel in dem Synset vorkommen. Werden Hyponym und Hyperonym mit einbezogen, wird das Beispiel dagegen korrekt klassifiziert. Bei dem folgenden Synset-Artikel Paar findet ebenso eine korrekte Klassifikation statt, obwohl die Anzahl der übereinstimmenden Wörter nicht größer ist, allerdings ist der erste Absatz des Wikipedia Artikels dort deutlich kleiner:

- Wikipedia: A surface lift is a mechanical system used to transport skiers and snowboarders where riders remain on the ground as they are pulled uphill.
- WordNet: surface lift (a ski tow that pulls skiers up a slope without lifting them off the ground)

Das bedeutet nicht, dass es sinnvoll wäre beispielsweise jeweils nur den ersten Satz eines Wikipedia Artikels in die Bag of Words aufzunehmen. Das Problem ist vielmehr, dass die betrachteten Bedeutungen in WordNet und Wikipedia unterschiedlich kompakt beschrieben werden.

5 Related Work

Das Mapping verschiedener Wissensbasen ist in der Forschung ein sehr aktuelles Thema zu dem momentan sehr viele Arbeiten veröffentlicht werden. Trotzdem fehlt es noch an öffentlich verfügbaren Datensätzen mit deren Hilfe erst ein wirklicher Vergleich auf Basis der Ergebnisse der veröffentlichten Ansätze möglich wird. Im Folgenden soll auf die wichtigsten thematisch eng verwandten Arbeiten eingegangen werden.

Suchanek *et al.* [2007] nutzen Informationen aus WordNet und Wikipedia zur Konstruktion einer Wissensbasis mit dem Namen „YAGO“ (Yet Another Great Ontology). Neben Bedeutungen wie sie aus WordNet bekannt sind, sollen auch Objekte wie Personen, Orte, Bücher usw., welche vor allem in Wikipedia zu finden sind, sowie zusätzliche Relationen als Fakten (z.B. „wer wurde wo geboren“ oder „wer hat was gewonnen“) angeboten werden. YAGO beschränkt sich dabei auf Nomen und ignoriert die in WordNet vorhandenen Verben, Adjektive und Adverbien. Statt einer wirklichen Synset-Artikel Zuweisung werden sämtliche WordNet Synsets und zusätzlich alle Wikipedia Artikel, deren Titel nicht mit einem WordNet Synset übereinstimmen, aufgenommen. Auf diese Weise gehen einige Bedeutungen wie der Artikel über die Rockband „Queen“ verloren, da es bereits ein WordNet Synset mit dem Synonym „queen“ gibt. YAGO hat anders als in dieser Arbeit nicht das Ziel eine allgemeine erweiterte Wissensbasis mit einem breiten Informationsangebot zu jedem Eintrag durch eine echte Verknüpfung der Ressourcen zu schaffen, sondern es werden Objekte gesammelt (insgesamt etwa 1 Million) und spezielle Fakten (etwa 5 Millionen) dazu aus den Quellen WordNet und Wikipedia zusammengetragen. Typische Anwendungen sind beispielsweise Anfragen wie „Welche Wissenschaftler wurden im Jahr 1879 geboren“. Derartige Fragen können jedoch nur für entsprechend vorhandene Relationen gestellt werden.

Des Weiteren gibt es aktuelle Arbeiten zur Zuweisung von WordNet Synsets und Wikipedia Kategorien. Besonders hervorzuheben ist dort die Arbeit von Toral *et al.* [2009], welche ebenso wie diese Arbeit die Zuweisung mit einem semantischen Textvergleich in einer Bag of Words Methode über mit Personalisiertem PageRank erstellte Vektoren durchführt. Als Ähnlichkeitsmaß wird die Cosinus Distanz verwendet. Die Bags of Words bestehen lediglich aus dem Gloss des betrachteten Synsets bzw. aus dem Abstract der betrachteten Kategorie. Zu beachten ist, dass keine Wikipedia Artikel, sondern nur Wikipedia Kategorien zugewiesen werden, deren Anzahl weit unterhalb der Anzahl an Wikipedia Artikeln liegt (etwa 3.4 Millionen Artikel gegenüber 0.5 Millionen Kategorien). Die Ergebnisse lassen sich daher nicht direkt vergleichen. Als Baseline wurde ein auf Wortüberlappung basierendes Verfahren genutzt (ähnlich der hier verwendeten String-basierten Methode), sowie eine First-Sense Heuristik (einer Wikipedia Kategorie wird aus allen das Synonym enthaltenden Synsets immer das allgemein am häufigsten verwendete Synset zugewiesen). Die Evaluation zeigt, dass das PageRank-basierte Verfahren auch hier das auf Wortüberlappung beruhende Verfahren übertrifft. Neben dem PageRank-basierten Verfahren werden weitere grundlegend verschiedene Verfahren getestet, auf die hier jedoch nicht näher eingegangen werden soll. Das PageRank-basierte Verfahren erreicht im Vergleich zu den anderen getesteten unüberwachten Verfahren gute Werte, wenn auch leicht unterhalb der First-Sense Heuristik Baseline. Festzustellen ist außerdem, dass sich das Entfernen von StopWords aus Gloss bzw. Kategorie Text merkbar positiv auf die Ergebnisse auswirkt. Zusätzlich werden Kombinationen der Verfahren getestet, was zu einer Verbesserung der Ergebnisse führt. Eine Übertragung entsprechender Ansätze auf die Artikel-Ebene und eine Kombination mit dem hier vorgestellten Verfahren wäre denkbar.

Ebenfalls mit einer Zuweisung von WordNet Synsets und Wikipedia Kategorien beschäftigt sich die Arbeit von Ponzetto and Navigli [2009]. Der Ansatz versucht zunächst vollständig übereinstimmende Kategori-

en (Titel) und Synsets (Synonyme) einander zuzuweisen. Des Weiteren werden Wikipedia Kategoriebäume und die WordNet Hyponym/Hyperonym Hierarchie verwendet, um jeder Kategorie das relevanteste Synset zuzuweisen. Dabei haben lediglich die Kategorie Namen und Synset Synonyme Einfluss auf die Zuweisung. Letzlich handelt es sich bei dem Algorithmus ebenso um ein erweitertes String-basiertes Verfahren.

Neben der vereinfachten Synset-Kategorie Zuweisung existieren auch Ansätze, die sich mit einer Synset-Artikel Zuweisung auseinandersetzen. Ruiz-Casado *et al.* [2005] präsentieren einen solchen Ansatz, der allerdings nur auf „Simple Wikipedia“¹ angewandt wird. Simple Wikipedia ist eine Version von Wikipedia, die nur einfache englische Wörter nutzt und das Ziel hat für Menschen mit eingeschränkten Englischkenntnissen möglichst verständlich zu sein. So besitzt Simple Wikipedia lediglich 64.682 Artikel², was gegenüber der echten englischsprachigen Wikipedia mit über 3,4 Millionen Artikeln verschwindend wenig ist, sodass auch hier kein direkter Vergleich möglich ist. Die Anzahl der Artikel ist sogar deutlich unterhalb der Anzahl der in WordNet vertretenen Synsets. In dem verwendeten Verfahren wird zunächst für jeden Artikel Titel geprüft, ob er in genau einem WordNet Synset als Synonym vorkommt und in diesem Fall diesem zugewiesen. Ist er in keinem Synset enthalten, wird er ignoriert, falls er in gleich mehreren Synsets enthalten ist, wird ein auf Wortüberlappung basierendes Verfahren eingesetzt. Einbezogen werden in diesem Verfahren neben dem Artikel Text die Glosswörter des jeweiligen Synsets, sowie Synonyme von Hyponymen und Hyperonymen. Die Ähnlichkeiten werden mit Punktprodukt und Cosinus Distanz berechnet. Es wird dann grundsätzlich das Synset mit der höchsten Ähnlichkeit dem Artikel zugewiesen.

Erst kürzlich veröffentlicht wurden zudem zwei Arbeiten, die sich tatsächlich mit einem Mapping von WordNet und Artikeln der vollständigen englischen Wikipedia auseinandersetzen. Ponzetto and Navigli [2010] beispielsweise stellen einen Algorithmus vor, der basierend auf einem String-basierten Überlappungs-Verfahren Zuweisungen vornimmt. Dazu werden zunächst (ähnlich wie in dieser Arbeit) Bags of Words aufgebaut: Als Kontext für Wikipedia dient der Artikel Titel, die ausgehenden Verweise des Artikels, sowie die dem Artikel zugeordneten Kategorien. Aus WordNet werden Gloss und Synonyme des betrachteten Synsets aufgenommen, sowie Synonyme der Hyperonyme, Hyponyme und Geschwister (zwei Synsets sind Geschwister, wenn sie ein gemeinsames Hyperonym haben). Der Zuweisungs-Algorithmus geht so vor, dass zunächst Paare einander zugewiesen werden, deren Artikel Titel und Synset Synonym vollständig übereinstimmen. Andernfalls wird ein auf Wortüberlappung basierendes Verfahren auf die Bags of Words angewandt. Es handelt sich letztlich also um einen optimierten String-basierten Ansatz. Dieser wurde auf einem Gold-Standard³ getestet, der erst kürzlich veröffentlicht wurde, sodass es nicht mehr möglich war, die hier vorgestellte PageRank-basierte Methode auf diesem Gold-Standard auszuführen um die Ergebnisse direkt vergleichen zu können. Interessant an der Arbeit von Ponzetto and Navigli [2010] ist jedoch vor allem, dass sie nicht in der Theorie verbleibt, sondern mit der vorgestellten Methode tatsächlich eine erweiterte Wissensbasis „WordNet++“ geschaffen wird. Diese wird mit existierenden Word Sense Disambiguation Algorithmen getestet, welche „WordNet++“ als Wissensbasis nutzen. Im Ergebnis zeigen sich deutliche Verbesserungen gegenüber der einfachen Nutzung von WordNet oder Wikipedia als Wissensbasen. Damit ist ein Nachweis für die Motivation dieser Arbeit erbracht.

Eine weitere sehr aktuelle Arbeit von de Melo and Weikum [2010] hat das Ziel aus WordNet eine multilinguale Wissensbasis zu erzeugen, indem WordNet Synsets und englischsprachige Wikipedia Artikel verknüpft werden. Die meisten Artikel enthalten Verweise auf entsprechenden Artikel in anderen Sprachen, aus denen dann neue Synsets anderer Sprachen generiert werden können (Titel und Redirects entsprechen Synonymen, der Gloss wird aus dem ersten Absatz des jeweiligen Artikels gewonnen). Für

¹ <http://simple.wikipedia.org>

² Stand: September 2010

³ <http://lcl.uniroma1.it/babelnet/>

die Mappings zwischen WordNet und Wikipedia werden drei verschiedene Heuristiken gemittelt. Dabei wird zunächst die Überlappung von Synonymen mit Titel und Redirects untersucht. Das zweite Maß berechnet die Ähnlichkeit von Vektoren bestehend aus Gloss bzw. erstem Artikel Absatz mit der Cosinus-Distanz (ähnlich der hier vorgestellten String-basierten Methode). Als drittes Maß fließt schließlich eine First Sense Heuristik ein. Auch hier kann folglich nicht von einem semantischen Ansatz gesprochen werden, wie ihn die PageRank-basierte Methode darstellt.

Abschließend noch ein kurzer Blick auf die Arbeit von Fernando and Stevenson [2010], die sich ausschließlich mit der Vorauswahl von Wikipedia Artikeln für WordNet Synsets beschäftigt. Dies ist insofern von großer Relevanz, da eine gute Vorauswahl die spätere Zuweisung vereinfacht. In dem Paper werden verschiedene Vorgehensweisen getestet. Eine Methode ist beispielsweise alle Artikel für ein Synset auszuwählen, deren Titel oder Redirect einem der Synonyme entsprechen. Zusätzlich können die in den ausgewählten Artikeln zugehörigen Disambiguierungs-Seiten enthaltenen Verweise auf weitere Artikel genutzt werden. Der Recall Wert steigt dabei mit der Anzahl zugelassener möglicher Artikel für ein Synset. Mit einer Kombination der vorgestellten Methoden ist auf dem Testdatensatz bei einer Beschränkung auf 10 Artikel ein Recall Wert von 0.931 möglich (was allerdings keine allzugroße Erleichterung der im Anschluss angewandten Zuweisungsalgorithmen darstellt).

Alles in allem ist zusammenzufassen, dass bislang nach unserem Wissen keine Arbeit existiert, die sich mit einem vollständigen Mapping von WordNet und Wikipedia Artikeln beschäftigt und dabei auf einem semantischen Textvergleich basiert, wie er in dieser Arbeit vorgestellt wurde. Viele der vorgestellten Verfahren haben eine große Ähnlichkeit mit der in dieser Arbeit als Baseline verwendeten String-basierten Methode, wobei diese meistens optimiert wurde. So wird beispielsweise des Öfteren versucht zunächst Artikel und Synsets direkt einander zuzuweisen, wenn Synonyme und Artikel Titel übereinstimmen, worin gewisse Parallelen zu dem von uns verwendeten Bonussystem zu erkennen sind. Für die Zukunft wären mehr öffentlich verfügbare Testdatensätze wünschenswert, um die in den verschiedenen Arbeiten gewonnenen Erkenntnisse besser einordnen zu können.

6 Zusammenfassung

In dieser Bachelorarbeit wurde ein semantischer Ansatz für die Zuweisung von WordNet Synsets und Wikipedia Artikeln vorgestellt (Abschnitt 2.2). Dieser basiert auf einer Bag of Words Methode und dem aus der Word Sense Disambiguation bekannten Personalisierten PageRank Verfahren von Agirre and So-roa [2009]. Ziel der Arbeit ist es, über die Lösung dieses Zuweisungsproblems die Voraussetzung für die automatische Konstruktion einer erweiterten Wissensbasis aus WordNet und Wikipedia zu schaffen, welche eine höhere Abdeckung von Bedeutungen, bessere semantische Beziehungen, sowie eine größere Menge an Informationen enthält und somit einen Beitrag zur Verbesserung existierender Anwendungen der natürlichen Sprachverarbeitung wie Word Sense Disambiguation leisten kann. Als Baseline für die Bewertung unseres semantischen Ansatzes haben wir ein String-basiertes Verfahren genutzt (Abschnitt 2.1), das ausschließlich auf der Überschneidung von Wörtern basiert, in ähnlicher Form jedoch von vielen aktuellen Arbeiten genutzt wird (siehe Kapitel 5).

Die Evaluation (Kapitel 4) hat gezeigt, dass dieser neuwertige semantische Ansatz alleine die Ergebnisse der als Baseline verwendeten rein String-basierten Methode deutlich übersteigt. Auf dem verwendeten Gold-Standard sind mit dem PageRank-basierten Verfahren F-Measure Werte bis 0.776 möglich, während die String-basierte Methode nicht über 0.738 kommt. Für die PageRank-basierte Methode wurden insgesamt fünf verschiedene Ähnlichkeitsmaße getestet (Abschnitt 2.3), deren Ergebnisse sich deutlich voneinander unterscheiden. Das am besten für diesen Zweck geeignete Maß ist χ^2 . Mit der recht einfachen Kombinationsmethode (Abschnitt 2.4) sind sogar Werte bis 0.781 möglich. Bei einer Auswertung mit Cross-Validation variieren die Ergebnisse nur leicht, was auf eine hohe Stabilität der Verfahren auch auf größeren Datensätzen schließen lässt. Außerdem wurde ein Bonussystem getestet, mit dem in der Kombination auf den Trainingsdaten Werte bis zu 0.799 erreicht wurden. Die Unterschiede zwischen der String-basierten Methode und der PageRank-basierten Methode alleine waren mit Bonus nicht mehr so eindeutig, was daran liegt, dass insbesondere bei der String-basierten Methode das Bonussystem die Ergebnisse sehr stark beeinflusst. Bei der Auswertung wurden verschiedene Zusammensetzungen der Bags of Words getestet. Dabei ergaben sich in den Ergebnissen deutliche Unterschiede. Für WordNet hat sich herausgestellt, dass insbesondere die Hinzunahme von Hyperonymen in die Bag of Words von Vorteil ist, für Wikipedia sind dies die Kategorien des jeweiligen Artikels. Weitere Verbesserungen des vorgestellten PageRank-basierten Ansatzes wurden in Abschnitt 4.2 diskutiert.

Ein direkter Vergleich der Ergebnisse mit anderen Arbeiten zu der Thematik war zum Zeitpunkt der Fertigstellung dieser Arbeit aufgrund fehlender öffentlicher Datensätze nicht möglich. Es ist jedoch festzustellen, dass keine Arbeit existiert, die sich mit einer Zuweisung von WordNet Synsets und Artikeln der gesamten englischsprachigen Wikipedia auseinandersetzt und dabei auf einen semantischen Ansatz wie die PageRank-basierte Methode setzt.

Abbildungsverzeichnis

1.1	Semantische Relationen zwischen Synsets in WordNet	6
1.2	Beispiel einer Disambiguierungsseite von Wikipedia	7
1.3	Erster Absatz eines Wikipedia Artikels	8
1.4	Erstellung zweier Bags of Words aus einem WordNet Synset (links) und dem ersten Absatz eines Wikipedia Artikels (rechts)	9
2.1	Die Vorgehensweise im Überblick	10
2.2	Erstellung von Vektoren aus Bag of Words	10
2.3	PageRank	11
2.4	Einige PageRank Iterationen ($c=0.5$): Die Werte in den Knoten beschreiben die aktuellen Gewichte, die Kanten bilden die Gewichte ab, die an andere Knoten abgegeben werden. Der letzte Graph zeigt Konvergenz	12
2.5	Beispiel für WordNet Graph mit zwei verschiedenen Bedeutungen für das Wort „Bank“	13
2.6	Vorgehensweise von Subgraph- und Personalisierter PageRank Methode (bei WSD)	14
2.7	Erstellung repräsentativer Vektoren mit Personalisiertem PageRank	14
2.8	Cosinus Distanz	16
2.9	Beispiel von zu vergleichenden Vektoren (String-basierter Ansatz)	16
3.1	Preprocessing Pipeline	18
3.2	Ausschnitt aus dem Gold-Standard	19
3.3	Bag of Words zum Wikipedia Artikel „Actor“	19
3.4	Training/Testing Pipeline (* Annotator ist je nach Aufbau der Pipeline optional)	20
3.5	Beispiel für vom Evaluator ausgegebene Kurven: Schwellenwertkurve (oben) und ROC-Kurve (unten), wobei $FPR = \text{False Positive Rate} = FP/(FP+TN)$ und $TPR = \text{True Positive Rate} = TP/(TP+FN)$	22
3.6	Beispiel für vom Evaluator ausgegebene Recall-Precision Kurve	23
3.7	Beispiel für ein Ausgabefile des CrossValidators	23
4.1	Beschränkung der Anzahl einem Synset zugewiesener Artikel	26

Tabellenverzeichnis

4.1	Ergebnisse der automatischen Zuweisung ohne Bonussystem auf Trainingsdaten (T) und mit Cross Valdiation (CV), maximal ein Artikel pro Synset (Werte in F-Measure)	31
4.2	Ergebnisse der automatischen Zuweisung mit Bonussystem auf Trainingsdaten (T) und mit Cross-Validation (CV), maximal ein Artikel pro Synset (Werte in F-Measure)	32

Literaturverzeichnis

- [Agirre and Soroa, 2008] E. Agirre and A. Soroa. Using the multilingual central repository for graph-based word sense disambiguation. In *Proceedings of the Conference on Language Resources and Evaluation (LREC'08)*, Marrakesh, Morocco, 2008.
- [Agirre and Soroa, 2009] E. Agirre and A. Soroa. Personalizing PageRank for WordSenseDisambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL)*, pages 33–41, Athens, Greece, 2009.
- [Agirre et al., 2009] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A study on similarity and relatedness using distributional and WordNetbased approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 19–27, Boulder, USA, 2009.
- [Banerjee and Pedersen, 2003] S. Banerjee and T. Pedersen. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.
- [Brin and Page, 1998] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 30(1–7), 1998.
- [de Melo and Weikum, 2010] G. de Melo and G. Weikum. Providing Multilingual, Multimodal Answers to Lexical Database Queries. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.
- [Fellbaum, 1998] C. Fellbaum. WordNet: An electronic lexical database. The MIT Press, 1998.
- [Fernando and Stevenson, 2010] S. Fernando and M. Stevenson. Aligning WordNet Synsets and Wikipedia Articles. In *Workshop on Collaboratively built Knowledge Sources and Artificial Intelligence, AAAI-2010*, 2010.
- [Gabrilovich and Markovitch, 2006] E. Gabrilovich and S. Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, Boston, Massachusetts, USA, 2006.
- [Giles, 2005] J. Giles. Internet encyclopaedias go head to head. In *Nature*, pages 900–901. Nature Publishing Group, 2005.
- [Glickman et al., 2005] O. Glickman, I. Dagan, and M. Koppel. Web based probabilistic textual entailment. In *PASCAL Challenges Workshop on RTE*, 2005.
- [Mihalcea and Moldovan, 2001] R. Mihalcea and D.I. Moldovan. eXtended WordNet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Ressources*, pages 95–100, 2001.
- [Mihalcea et al., 2006] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2006.

-
- [Mihalcea, 2005] R. Mihalcea. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT05)*, pages 411–418, Morristown, NJ, USA, 2005.
- [Mihalcea, 2007] R. Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA, 2007.
- [Navigli and Lapata, 2007] R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.
- [Ponzetto and Navigli, 2009] S.P. Ponzetto and R. Navigli. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21th International Joint Conference on Artificial Intelligence (IJCAI'09)*, pages 2083–2088, 2009.
- [Ponzetto and Navigli, 2010] S.P. Ponzetto and R. Navigli. Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 1522–1531, 2010.
- [Ruiz-Casado *et al.*, 2005] M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Advances in Web Intelligence*, pages 380–386. Springer, 2005.
- [Suchanek *et al.*, 2007] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW'07)*, pages 697–706, 2007.
- [Toral *et al.*, 2009] A. Toral, O. Ferrandez, E. Agirre, and R. Munoz. A study on Linking Wikipedia categories to Wordnet using text similarity. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2009.
- [Zesch *et al.*, 2007] T. Zesch, I. Gurevych, and M. Muehlhaeuser. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 205–208, 2007.
- [Zesch *et al.*, 2008] T. Zesch, C. Mueller, and I. Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC'08)*, pages 1646–1652, 2008.