IMS

Bachelorarbeit Nr. 207

Wissensbasierte lexikalische Substitution

Ilhan Tas

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Sebastian Padó
Betreuer/in:	Prof. Dr. Sebastian Padó
Beginn am:	26. März 2015
Beendet am:	16. September 2015
CR-Nummer:	H.3.1, H.3.3, I.2.7

Kurzfassung

Lexikalische Mehrdeutigkeit ist eine fundamentale Eigenschaft von Sprachen, in denen viele Wörter mehrere sich von einander unterscheidende Bedeutungen haben. Wohingegen eine Person beim Lesen eines Textes oder in der Führung einer Konversation ihr angeeignetes Wissen beziehungsweise die Lebenserfahrung und den Kontext zu Hilfe nimmt um die richtige Lesart zu bestimmen, sieht dieser Prozess bei einem Rechner anders aus. Für diesen sind Texte nichts anderes als Zeichenketten respektive eine Aneinanderreihung von Buchstaben. Folglich müssen die ambigen Wörter, im Hinblick auf einen maschinellen Umgang mit natürlicher Sprache, aufgelöst und die richtige Lesart bestimmt werden. Lexikalische Mehrdeutigkeit ist ein weitreichendes Problem der maschinellen Verarbeitung natürlicher Sprache und gehört zu den immer mehr an Bedeutung gewinnenden Forschungsgebieten der Computerlinguistik.
Lexikalische Substitution ist ein relativ neues Paradigma zur Lösung dieses Problems und wurde von McCarthy und Navigli bei SemEval 2007 eingeführt. Ziel dieser Herangehensweise ist die Generierung und das Ranking von Substitutionskandidaten für ein Zielwort im Hinblick auf ihre Angemessenheit bezüglich des Kontexts, in dem das zu ersetzende Wort erscheint. Dieser Ansatz ist eng verwandt mit der Wortbedeutungsdisambiguierung (engl. Word Sense Disambiguation, kurz WSD). Die drei Hauptansätze in der lexikalischen Substitution lassen sich in die Kategorien überwachte, unüberwachte und wissensbasierte Systeme eingliedern.
Diese Arbeit stellt ein wissensbasiertes Modell für ein vokabular-globales Substitutionssystem vor. Grundlage hierfür ist der Lesk-Algorithmus. Mithilfe des Lesk-Algorithmus' wird ein Ranking für potentielle Substitute aufgestellt und anschließend lexikalische Substitution auf dem CoInCo-Korpus durchgeführt.

Abstract

Lexical ambiguity is a fundamental characteristic of languages, in which many words have multiple and different meanings. While a person can use his experiences or the context to identify the intendent meaning of an ambiguous word in a text or a conversation, this process still looks different in computers. For computers texts are nothing else than strings respectively a sequence of characters. Hence with regard to a mechanical handling of natural languages, ambiguous words must be disambiguated to find the correct reading in a specific context of use. Lexical ambiguity is a far-reaching problem of natural language processing and is one of increasingly important research areas of computational linguistics.

Lexical substitution is a novel approach to deal with ambiguity and was introduced during SemEval 2007 by McCarthy and Navigli. The aim of this approach is the identification and ranking of substitutions for a target word in terms of their appropriateness with regard to the context in which the word to be replaced appears. This task is strongly related to Word Sense Disambiguation (WSD). The three main approaches of lexical substitution can be monitored to the categories supervised, unsupervised and knowledge-based systems.

This thesis presents a knowledge-based vocabulary-global substitution system. Based on the Lesk-Algorithm a ranking is done for candidate substitutes and finally lexical substitution is performed on the CoInCo corpus.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung und Motivation

Die Kommunikation über Sprache gehört zum festen Bestandteil des menschlichen Alltags. Diese erstaunliche Fähigkeit ermöglicht Menschen sich auszutauschen, Information zu beziehen und zu verbreiten. Sprache als Medium funktioniert, weil ihre Benutzer über ein Weltwissen verfügen, welches es ermöglicht den wahrgenommen Lauten eine Bedeutung zuzuordnen.
Bedingt durch den wissenschaftlichen und technischen Fortschritt ist dieses Medium kein exklusives innermenschliches Phänomen mehr. Heute findet Kommunikation auch zwischen Menschen und Maschinen statt. Die Interaktion über Sprache mit dem Smartphone, die Steuerung des Radios oder Navigationssystems im Auto und die Kommunikation mit Suchmaschinen wie Google sind nur einige Beispiele hierfür. Dieser Art von Kommunikation sind jedoch Grenzen gesetzt, da den Maschinen das nötige Weltwissen fehlt.
Viele Wörter natürlichsprachlicher Texte weisen verschiedene Lesarten auf und sind folglich mit Mehrdeutigkeit behaftet (Kremer et al., 2014). Th. Lewandowski (Lewandowski, 1994) beschreibt den Bergriff Lesart folgendermaßen:

Definition 1.0.1 (Lesart)
"[reading]. Eine bestimmte semantische Repräsentation eines Morphems, Wortes oder Satzes".

Folglich versteht man unter einer Lesart eine mögliche Bedeutungsvariante eines mehrdeutigen Wortes. Der Fachausdruck für dieses Phänomen von Wörtern mit mehreren Lesarten lautet in der Fachliteratur Ambiguität. Eine Definition für Ambiguität findet sich im Lexikon der Sprachwissenschaft (Bußmann, 2008).

Definition 1.0.2 (Ambiguität)
"[lat. ambiguitās ›Doppelsinn‹. - Auch: Amphibolie (veraltet), Mehrdeutigkeit] Eigenschaft von Ausdrücken natürlicher Sprachen, denen mehrere Bedeutungen zukommen. Ambige Ausdrücke sind (isoliert betrachtet) semantisch unbestimmt und folglich präzisierungsbedürftig. Ambiguität zeichnet sich dabei gegenüber Vagheit dadurch aus, dass das Präzisierungsspektrum als diskret wahrgenommen wird (Bsp. Bank: Lesart 1 = >Geldinstitut<, Lesart 2 = >Sitzgelegenheit< usw.), während vage Ausdrücke (z.B.Farbadjektive, Gradadjektive) über ein kontinuierliches Präzisierungsspektrum verfügen."

Neben ambigen Ausdrücken, von denen in Definition 1.0.2 die Rede ist, können auch Paraphrasen und sogar ganze Sätze mit einer Bedeutungsvielfalt versehen sein. Diese Eigenschaft von (primär) Wörtern ist der ökonomischen Tendenz von natürlichen Sprachen geschuldet, die mit einem begrenzten Wortinventar etwas zu beschreiben Versuchen, dessen Grenzen unendlich sind.

Das folgende Zitat von Weaver (1949) verdeutlicht das Problem sowie den Lösungsansatz:

„If one examines the words in a book, one at time through an opaque mask with a whole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. "Fast"may mean "rapid"; or it may mean "motionless"; and there is no way of telling which. But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguosly decide the meaning“ (Weaver, 1949).

Fasst man die Kernaussage des Zitats zusammen, so lautet sie, dass Wörter alleinstehend ambig sind und es unmöglich zu bestimmen ist, welche ihrer Bedeutungsvarianten sie repräsentieren. Betrachtet man sie jedoch in einem ausreichend großen Kontext, also etwa in einem Satz oder einem Abschnitt, kann man entscheiden welche Bedeutung sie annehmen.

Ambiguität ist ein sehr weit verbreitetes Problem und kommt in allen natürlichen Sprachen vor. Betrachtet man die 121 am häufigsten in realen Texten vorkommenden Nomen und die häufigsten 70 Verben der englischen Sprache, so haben die Nomen im Schnitt ca. 8 und die Verben sogar 12 verschiedene Bedeutungen[1] (Miller et al., 1990), (Ng und Lee, 1996).

Die Ambiguität bereitet Probleme bei der automatischen Datenextraktion aus Texten. Insbesondere in den Bereichen Wörterbucherzeugung, Grammatikerstellung sowie Information Retrieval gestaltet sich die Mehrdeutigkeit als eine Fehlerquelle und muss aufgelöst werden. Das Wort *Bank* kann beispielsweise ein Geldinstitut oder ein Sitzmöbelstück darstellen. Der traditionelle Ansatz im Umgang mit der lexikalischen Mehrdeutigkeit liegt in der überwachten Wortbedeutungsdisambiguierung (engl. Word Sense Disambiguation, kurz WSD). Ziel dieses Ansatzes ist die Identifizierung der intendierten Bedeutung eines Wortes unter Zuhilfenahme der Kontextinformation. Neben der Kontextinformation sind semantisch angereicherte Ressourcen für die Disambiguierung unerlässlich. Diese sind zum Beispiel Datenbanken wie WordNet[2], welche semantische und lexikalische Beziehungen zwischen den Wörtern enthalten und insbesondere ein Lesarteninventar zur Verfügung stellen (Fellbaum, 1998). Ein Disambiguierungssystem weist nun mit Hilfe des Lesarteninventars jedem Vorkommen eines ambigen Zielworts den passenden Lesartenindex zu (Carstensen et al., 2010). Folglich handelt es sich bei der Lesartendisambiguierung mit der oben beschriebenen Methode um ein Klassifikationsproblem.

Lexikalische Substitution ist ein relativ neues Paradigma zur Bedeutungsbeschreibung, das eine Alternative zu Word Sense Disambiguation darstellt. Instanzenbedeutung wird hier nicht per Referenz auf eine Liste von Synonymen (engl. synsets) repräsentiert, sondern als Menge an möglichen Substituten (McCarthy und Navigli, 2009). Betrachtet man zum Beispiel[3] den Satz *"He was **bright** and independent and proud"*, in dem das Adjektiv ***bright*** ersetzt werden soll, so wären *intelligent*, *clever* oder *smart* angemessene Substitute.
Lexikalische Substitution ist eng verwandt mit der Wortbedeutungsdisambiguierung. Beide Ansätze beschäftigen sich mit der Identifizierung von Substituten für ein Zielwort, unter Zuhilfenahme des Kontexts, ohne dabei die ursprüngliche Bedeutung des Satzes zu ändern. Der Unterschied der beiden Ansätze liegt darin, dass die lexikalische Substitution das Inventar an Substituten beziehungsweise

[1] Unter der Voraussetzung von WordNet 1.5 als Lesarteninventar

[2] WordNet ist eine lexikalische Datenbank für das Englische. Aufgrund ihrer freien Verfügbarkeit wird sie oft in Disambiguierungsprozessen eingesetzt. Verfügbar auf: https://wordnet.princeton.edu/wordnet/download/

[3] Dieses Beispiel stammt aus dem LexSub Datensatz. Die Substitute *intelligent* und *clever* wurden von je 3 Annotatoren und das Substitut *smart* von einem Annotator vorgeschlagen.

deren Quelle nicht explizit vorschreibt und dadurch das Problem der Granularität der Bedeutungsdifferenzierung überwindet (McCarthy und Navigli, 2009), (McCarthy, 2002). Die Hauptaufgaben bei der lexikalischen Substitution sind die Generierung und das Ranking der Substitute. Ersteres befasst sich dabei mit der Erzeugung beziehungsweise Erarbeitung von möglichen Synonymen und letzteres beschäftigt sich mit der Bewertung der erarbeiteten Substitute in Bezug auf den Kontext und erstellt eine Rangliste bezüglich der Güte der Substitute für den jeweiligen Kontext (Szarvas et al., 2013).

Die Zielsetzung dieser Arbeit liegt in der Durchführung der lexikalischen Substitution. Hierfür wird ein wissensbasierter, alle Wörter umfassender, Ansatz verfolgt, welcher auf dem vereinfachten Lesk-Algorithmus basiert.

Gliederung

Die Arbeit ist in folgender Weise gegliedert:

Kapitel 2 – Eine Einführung in den Begriff der Ambiguität: In diesem Kapitel werden die grundlegenden Begriffe erläutert. Insbesondere wird auf den Begriff *Ambiguität* in der Semantik eingegangen.

Kapitel 3 – Stand der Forschung: Dieses Kapitel bietet Einblicke in die aktuelle Forschung in der lexikalischen Substitution und beschreibt einige Ansätze, die mit dieser Ausarbeitung thematisch verwandt sind.

Kapitel 4 – Datensätze und Ressourcen: Hier werden die verwendeten Datensätze und Ressourcen aufgelistet und es wird auf ihre Zusammensetzung sowie Struktur eingegangen.

Kapitel 5 – Eigener Ansatz: Der dieser Arbeit zugrunde liegende Ansatz wird in diesem Kapitel erläutert. Ferner wird auf die Methodik eingegangen, welche verfolgt wird um lexikalische Substitution durchzuführen.

Kapitel 6 – Evaluation: Im Evaluierungskapitel werden die durchgeführten Experimente, sowie die erzielten Ergebnisse wiedergegeben.

Kapitel 7 – Zusammenfassung und Ausblick: Das letzte Kapitel dieser Ausarbeitung fasst die Arbeit zusammen und diskutiert sie. Des weiteren bietet es Vorschläge zur Verbesserung der erzielten Ergebnisse.

2 Eine Einführung in den Begriff der Ambiguität

> You shall know a word by the company it keeps.
>
> *(John Rupert Firth)*

Für ein besseres Verständnis der Thematik werden in diesem Kapitel die dem Thema der Ausarbeitung zugrunde liegenden Begriffe eingeführt und anhand von Beispielen erläutert. Diese sind der Linguistik beziehungsweise der maschinellen Sprachverarbeitung und der Computerlinguistik zugeordnet.

2.1 Ambiguität

Im einleitenden Kapitel wurde bereits eine Definition für Ambiguität eingeführt. Sie beschreibt das Phänomen, dass nahezu alle Inhaltswörter mehrere Bedeutungen haben und sogar viele Sätze mehr als eine Lesart aufweisen können. Ambiguität liegt also immer dann vor, wenn ein Ausdruck beziehungsweise eine Äußerung nicht eindeutig interpretiert werden kann. Das Wort *Maus* kann zum Beispiel ein kleines Nagetier darstellen oder auch ein elektronisches Gerät um den Cursor auf dem Monitor zu steuern (Auberle et al., 2009). Um das Phänomen der Ambiguität zu konkretisieren, werden im Folgenden die linguistischen Grundlagen und Begriffe erläutert.

2.1.1 Lexem

Lexikalische Bedeutung wohnt nicht nur einzelnen Wörtern inne, sondern auch zusammengesetzten oder komplexen Ausdrücken. Komplexe Ausdrücke beziehungsweise Wortverbindungen sind solche, deren Gesamtbedeutung sich nicht aus den Einzelbedeutungen seiner Teile ableiten lässt und haben eine besondere Bedeutung. Beispiele hierfür sind *das Handtuch werfen, grauer Star* oder *ins Grass beißen*. Ein Lexem ist dann folglich ein einfacher oder komplexer Ausdruck mit einer lexikalischen Bedeutung. Nach Löbner machen folgende Eigenschaften ein Lexem aus (Löbner, 2003) :

- **Lautform, gesprochene Form**
- **Schriftform, orthographische Form**
- **Grammatische Kategorie**
- **Inhärente grammatische Eigenschaften**
- **Grammatische Formen, insbesondere unregelmäßige Formen**

- **Lexikalische Bedeutung**

Ein wesentlicher Unterschied in einem dieser Eigenschaften ist hinreichend dafür von zwei Lexemen auszugehen. Üblich ist auch der Begriff Lexikoneinheit. Die Summe aller Lexeme bildet das Lexikon oder den Wortschatz einer Sprache (Löbner, 2003).

2.1.2 Lexikalische Ambiguität

Betrachtet man die Einträge eines Wörterbuchs, wird man kaum auf ein Wort und insbesondere Inhaltswort treffen, welches nicht mehrere Bedeutungen hat. Auch bei Lexemen muss dieses Spektrum an Bedeutungen berücksichtigt werden. Würde man die Restriktion einführen, dass ein Lexem nur eine Bedeutung haben darf, so würde die Anzahl der Einträge eines Wörterbuchs enorm anwachsen. Daher gehen wir von mehreren Lexemen aus, falls die Bedeutungsvarianten eines Wortes nicht miteinander zusammenhängen. In dem anderen Fall gehen wir von einem Lexem aus. Der Fachausdruck für den ersten Fall lautet Homonymie. Den letzteren Fall beschreibt man in der Literatur als Polysemie. Beispielsweise stell das Wort *Bank/Bänke* vs. *Bank/Banken* ein Homonym dar. Polysemie hingegen liegt zum Beispiel bei dem Lexem *Läufer*[1] vor, welches unter anderem auf einen Teppich, eine Schachfigur oder eine laufende Person referiert. Beide Phänomene stellen eine Form von semantischer lexikalischer Ambiguität dar. Diese liegt immer dann vor, wenn die Bedeutung eines Ausdrucks nicht eindeutig ist (Löbner, 2003). Es gibt jedoch auch Formen von lexikalischer Ambiguität, die nicht semantisch sondern syntaktisch sind. Ein Beispiel hierfür ist der Satz *"Der Mann sah die Frau mit dem Fernrohr"*. Hier ist nicht klar, ob der Mann oder die Frau das Fernrohr hat. Des weiteren existieren auch Formen von semantischer Ambiguität, die nicht lexikalisch sondern strukturell sind. Handelt es sich bei dem Satz *"Jeder Mann kennt eine Frau"* jeweils nur um eine und dieselbe Frau oder kennt jeder Mann nur irgendeine Frau? Diese Bachelorarbeit konzentriert sich auf die lexikalische Ambiguität.

Homonymie

Betrachtet man das Wort *Maus*, bietet der Duden[2] unter anderem die Bedeutungsvarianten *kleines Nagetier, elektrisches Gerät* oder *(salopp) Geld*. Da diese verschiedenen Bedeutungen, unter anderem die erste und die letztere, in keinem Zusammenhang zueinander stehen geht man von verschieden Lexemen aus. Dieser wesentliche Unterschied in der lexikalischen Bedeutung ist ein Indiz für die Annahme von mehreren Lexemen und ist ein Beispiel für Homonymie. Linguisten sprechen auch oft von Homonymie, falls Ausdrücke verschiedene Ursprünge haben. Stimmen zwei Ausdrücke in allen Lexemeigenschaften überein außer in der lexikalischen Bedeutung, so spricht man von totaler Homonymie sonst von partieller Homonymie. Ein partielles Homonym ist *Bank/Bänke* vs. *Bank/Banken*. Ferner unternimmt man die Unterscheidung zwischen Homographie und Homophonie. Homographe sind Homonyme, die dieselbe Schriftform haben, wohingegen Homophone solche mit derselben Lautform sind (Löbner, 2003).

[1] Laut Guiness-Buch der Rekorde 1997 das deutsche Wort mit den meisten Bedeutungen (24).
[2] http://www.duden.de/

Polysemie

Polysemie ist ubiquitär in natürlichen Sprachen, in denen der Großteil der Ausdrücke über ein Spektrum an miteinander zusammenhängenden Bedeutungen verfügt. Der polyseme Charakter von Lexemen ist das Ergebnis „einer natürlichen ökonomischen Tendenz von Sprachen“ (Löbner, 2003). Begründen lässt sich dieser Umstand damit, dass man versucht neue Phänomene beziehungsweise Dinge mit Ausdrücken zu benennen, die bereits eine ähnliche Bedeutung haben oder die neue Bedeutungsvariante wird unter dem bereits existierenden Lexem abgelegt. Die verschiedenen Bedeutungsvarianten von *Läufer*, die bereits erwähnt wurden, haben alle etwas mit *Überwindung von Distanz* zu tun. Diese Beziehung zwischen den verschiedenen Bedeutungsvarianten ist ein Indiz für Polysemie. Diese tritt in Sprachen viel häufiger auf als Homonymie. Da Wörter in verschiedenen Sprachen in der Regel nicht auf dieselbe Weise polysem beziehungsweise homonym sind, spielen diese Phänomene eine wichtige Rolle in der automatischen Textübersetzung. Für eine korrekte Übersetzung muss die intendierte Bedeutung ermittelt werden (Löbner, 2003).

Vagheit

Vagheit liegt immer dann vor, wenn ein Ausdruck flexibel genutzt werden kann. Das heißt seine Benutzung ist kontextabhängig beziehungsweise kann sich von Fall zu Fall aufgrund gewisser subjektiver Kriterien ändern. Rührend daher kann es vorkommen, dass verschiedene Subjekte andere Auffassungen haben können. Das Wort *Baby* zum Beispiel ist ein vages Konzept. Es ist schwierig eine Grenze dazwischen zu ziehen, wann man von einem *Baby* sprechen kann und wann nicht. Der Wert von vagen Ausdrücken variiert auf einer kontinuierlichen Skala. Ähnlich ist dies auch bei Farbadjektiven der Fall oder auch generell bei Adjektiven, die einen Komparativ besitzen (Löbner, 2003).

2.1.3 Kontextuelle Ambiguität

Kontextuelle Ambiguität entsteht, wenn die lexikalische Bedeutung von Wörtern eine Verschiebung erfährt. Hierbei ist die Verschiebung auf den Kontext zurückzuführen. Es entsteht eine neue Bedeutung, die mit der ursprünglichen eng verwandt ist. Die Verschiebung kann dabei metonymischer oder auch metaphorischer Natur sein.

Metonymie

Die Metonymie bezeichnet eine Verschiebung der begrifflichen Interpretation. Während bei Polysemie die Bedeutungsvarianten eines Wortes explizit oder implizit in einem Lexikon aufgelistet sind, trifft dies für die Metonymie nicht zu. Hier ist die intendierte Bedeutung nicht wörtlich zu verstehen. Es liegt jedoch in irgendeiner Form ein Bezug zur wörtlichen Bedeutung vor. Metonymie liegt beispielsweise vor, wenn Institutionen personalisiert werden oder Eigennamen für die Werke von Personen stehen. Betrachtet man den Satz *"Die Firma rief an"*, handelt es sich sicherlich um eine Person aus der Institution, welche den Anruf tätigt (Carstensen et al., 2010).

Metapher

Eine Metapher stellt eine nicht-wörtliche Rede dar, welche eine Ähnlichkeit zwischen der Ausdrucks- und der Äußerungsbedeutung herstellt. Die intendierte Bedeutung wird bei Metaphorik mittels Analogien aus anderen Bereichen beziehungsweise Konzepten verdeutlicht. Für eine vergebliche Aufgabe wird zum Beispiel die Metapher *"die Nadel im Heuhaufen suchen"* verwendet.

3 Stand der Forschung

Bisherige Arbeiten im Bereich der lexikalischen Substitution befassen sich entweder mit der Erzeugung und anschließendem Ranking der Substitute oder nehmen sich nur der schwierigeren Aufgabe, dem Ranking, an. Erstere sind daher auf jedes Wort anwendbar, da sie keine vorverarbeiteten Daten benötigen. Letztere nehmen die Menge der möglichen Substitute aus den Testdaten und befassen sich mit der Evaluation ihrer Ranking-Methoden (Szarvas et al., 2013).
Ansätze zur Auflösung von Mehrdeutigkeit im Bereich der lexikalischen Substitution lassen sich in folgende drei Hauptmethoden unterteilen:

- **Wörterbuch- und wissensbasierte Methoden**
- **Unüberwachte Methoden**
- **Überwachte Methoden**

Im Folgenden werden die Charakteristika der jeweiligen Ansätze erläutert und Beispiele für diese angeführt.

3.1 Wörterbuch- und wissensbasierte Methoden

Wörterbuch- beziehungsweise wissensbasierte Ansätze sind folgende, welche sich primär auf Wörterbücher, Thesauri oder lexikalische Wissensquellen stützen, ohne jedoch Gebrauch von Korpusevidenzen zu machen. Diese können in der Regel auf alle Wörter angewendet werden und bieten eine gute Überdeckung (Agirre und Edmonds, 2007).

3.1.1 Algorithmus von Lesk

Der wohl bekannteste Vertreter des wörterbuch- und wissensbasierten Ansatzes ist der Lesk-Algorithmus (Lesk, 1986). Als einer der ersten Algorithmen für die semantische Disambiguierung benötigt er neben einer Menge an Wörterbucheinträgen, eine für jede Bedeutung, lediglich Wissen über den unmittelbaren Kontext, in dem das zu substituierende Wort erscheint. In Pseudocode sieht der Algorithmus wie folgt aus:

Algorithmus von Lesk
für jede Bedeutung *i* von W_1, **für jede** Bedeutung *j* von W_2 **berechne** *Überlappung(i, j)*, die Anzahl der gemeinsamen Wörter in den Definitionen der Wortbedeutungen von *i* und *j* **finde** *i* und *j* mit maximaler *Überlappung(i, j)* **ordne** Bedeutung *i* zu W_1 und Bedeutung *j* zu W_2

Der ursprüngliche Lesk-Algorithmus führt die Disambiguierung von W_1, im Hinblick auf W_2, auf die Überlappung in ihren Bedeutungsdefinitionen in Wörterbüchern zurück. Die Überlappung ist dabei die Anzahl der gleichen Wörter, welche jeweils in den Bedeutungsdefinitionen der Zielwörter vorkommen. Der Algorithmus wird anhand eines aus Lesk (1986) entnommenen Beispiel erläutert. Die zu disambiguirenden Wörter sind *pine* und *cone*. Benötigt werden nun Wörterbucheinträge für diese zwei Wörter. Der *Oxford Advanced Lerner's Dictionary* (Hornby und Wehmeier, 1995) bietet hierfür folgende Einträge:

pine
1 seven kinds of evergreen tree with needle-shaped leaves
2 pine
3 waste away through sorrow or illness
4 pine for something, pine to do something

cone
1 solid body which narrows to a point
2 something of this shape, whether solid or hollow
3 fruit of certain evergreen trees (fir, pine)

Aus diesen Einträgen folgt die größte Überlappung beziehungsweise Schnittmenge in den Einträgen 1 für *pine* und 3 für *cone* unter allen möglichen Kombinationen der Bedeutungen für diese zwei Wörter. Die 3 Wörter *evergreen, tree* und *pine* bilden dabei die Menge der Überlappung. Da es sich bei der ersten Bedeutung von *pine* um einen Nadelbaum handelt und die dritte Definition von *cone* eine Frucht (Tannenzapfen) eines Baums darstellt, stimmt diese Zuordnung.

3.1.2 Wissensbasiertes System von Hassan et al. (2007)

Hassan et al. stellen ein wissensbasiertes lexikalisches Substitutionssystem (SubFinder) vor, welches mittels verschiedener Techniken und Ressourcen die wahrscheinlichsten Substitute für ein Wort in seinem Kontext anbietet (Hassan et al., 2007). Für die Erstellung der Substitutskandidaten werden in diesem Ansatz verschiedene Quellen herangezogen. Zu diesen gehören neben WordNet die Microsoft Encarta Enzyklopädie[1], Roget's Thesaurus[2] (Roget, 1911) und bilinguale Wörterbücher. In ihren Experimenten erzielen sie bessere Ergebnisse, wenn sie sich in den Ressourcen auf WordNet und Encarta beschränken. Nach der Generierung der Substitutskandidaten aus diesen beiden Quellen

[1]Diese Ressource wurde von Microsoft zwischen den Jahren 1993 und 2009 herausgegeben.
[2]Verfügbar auf: http://www.thesaurus.com/Roget-Alpha-Index.html

erstellen sie mittels verschiedener Methoden ein Ranking für die erzeugten Substitute. Hierzu gehören unter anderem *Lexical Baseline*, bei der Substitute, die in beiden Quellen vorkommen, höher gewichtet werden und *Most Common Sense*, bei welcher das erste Substitut im ersten Synset in WordNet am höchsten gerankt wird. Unter den Ansätzen die 2007 in der *Lexical Substitution Task* auf dem LexSub-Datensatz (siehe Kapitel 4.3) teilnehmen, erreichen sie den besten Wert in der Abdeckung der von den Annotatoren am meisten genannten Substitute.

3.2 Unüberwachte Methoden

Die unüberwachten Ansätze benötigen im Gegensatz zu den überwachten Ansätzen keine annotierten Trainingsdaten und brauchen weniger Rechenzeit und Leistung (Zhou und Han, 2005). Diese Methoden zur Lesartendisambiguierung vermeiden nahezu komplett die Nutzung externer Informationsquellen und arbeiten direkt auf rohen, unannotierten Korpora (Agirre und Edmonds, 2007).

3.2.1 Unüberwachtes Vektormodell von Thater et al. (2011)

Thater et al. präsentieren ein unüberwachtes Modell, welches die kontextuelle Wortbedeutung mit Vektoren darstellt (Thater et al., 2011). Die Vektoren werden dabei durch die Wörter, die im unmittelbaren syntaktischen Kontext des Zielworts stehen, modifiziert. Die Kontextualisierung eines Vektors erfolgt dabei durch die Neugewichtung seiner Komponenten basierend auf verteilter Information der Kontextwörter. Bei der Modellierung der Kontextualisierung als Modifikation des Zielvektors wird in diesem Ansatz, im Gegensatz zu früheren, dieser Prozess nicht nur auf Varianten der Vektorkomposition beschränkt. Zusätzlich zu diesen Operationen werden auch weitere betrachtet, welche individuelle Vektorkomponenten neu gewichten. Des weiteren erachten sie einerseits den verteilten Ähnlichkeitswert zwischen zwei Wörtern, die die Vektoreinträge darstellen, andererseits die aktuellen Kontextwörter in einer gegebenen syntaktischen Position, als die effektivste Basis für die Neugewichtung. Eine Evaluierung des Modells wird auf zwei Aufgaben durchgeführt. Diese sind Paraphrasenranking und Wortsinndisambiguierung, wobei ersteres sich mit dem Ranking von möglichen Substituten im Hinblick auf ihre Plausibilität für den Kontext befasst.

Das durch Thater et al. vorgeschlagene Modell für Wortbedeutungen erlaubt die Berechnung von Vektordarstellungen für die individuelle Nutzung von Wörtern, welche die spezifische Bedeutung eines Worts in seinem Kontext innerhalb eines Satzes charakterisiert. So muss zum Beispiel der Vektor für das Verb *charge* in dem Ausdruck *charge a tax* die monetäre Bedeutung von diesem reflektieren, wohingegen der Vektor für die Okkurrenz in dem Ausdruck *charge a battery* auf eine Bedeutung verweisen muss, welche etwas mit Stromzufuhr zu tun hat.

Abbildung 3.1 zeigt die graphischen Darstellungen für verschiedene Vektoren von *charge*. Der Basisvektor ist in (a) zu sehen. Die weiteren Darstellungen zeigen zwei kontextualisierte Vektoren für *charge* in dem Kontext *charge a tax*, wobei der Vektor in (b) durch eine strikte und der Vektor in (c) durch eine Methode basierend auf semantischer Ähnlichkeit gewonnen wurde. Die Dimensionen der Basis- und Kontextvektoren repräsentieren das gemeinsame Auftreten von Wörtern in einer spezifischen syntaktischen Beziehung. Die Kontextualisierungsoperation verstärkt diejenigen Dimensionen des Basisvektors, welche durch den Kontext der zu betrachtenden spezifischen Instanz

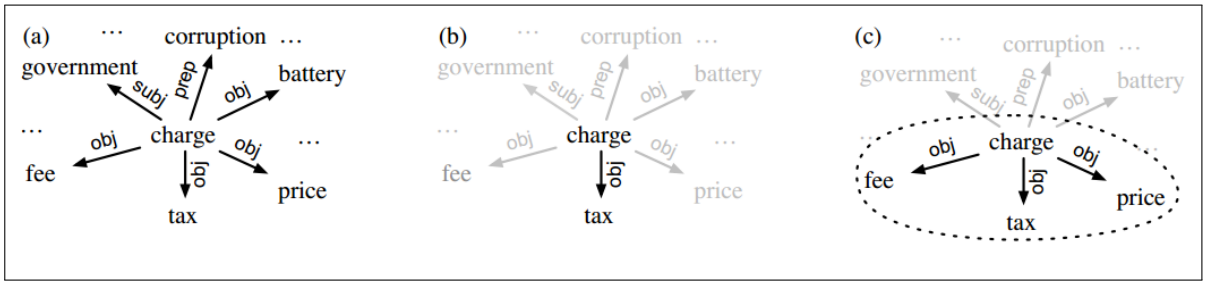

Abbildung 3.1: Basis- und Kontextvektoren für das Verb *charge*

zugelassen sind. Der einfachste Weg dabei, einen Wortvektor um seinen Kontext anzureichern besteht darin, nur die Dimensionen zu behalten, welche seinen syntaktischen Nachbarn entsprechen. Dieses Verfahren führt zu extrem dünn besiedelten Vektoren mit Nulleinträgen für die meisten Dimensionen. So führt zum Beispiel die Kontextualisierung des Vektors für *charge* in dem Kontext *charge a tax* (Abbildung 3.1b) dazu, dass für alle Tupel aus (Relation, Wort), bei denen $r \neq$ OBJ oder $w \neq$ *tax* ist, die Komponenten des Vektors den Wert Null erhalten. Die einzige Komponente ungleich Null bleibt die für *tax*. Auch wenn dieser einfache Ansatz überraschend erfolgversprechend ist, erzielen Thater et al. bessere Ergebnisse durch den Einsatz von Informationen aus der semantischen Ähnlichkeit von Kontextwörtern. Hierbei werden nicht nur die Dimensionen der Kontextwörter betrachtet, sondern zusätzlich auch die Dimensionen der Wörter, welche eine Verteilungsähnlichkeit zu diesen aufweisen. Dies resultiert dann in einem Vektor für *charge* (Abbildung 3.1c) in *charge a tax* der zusätzlich nicht Nulleinträge für die Dimensionen der Wörter enthält, welche ähnlich zu *tax* sind. Formal betrachtet wird in diesem Modell eine Menge W Wörter, sowie eine Menge R an syntaktischen Relationen angenommen. Die Relationen sind dabei zum Beispiel *Subjekt* oder *Objekt*. Die Bedeutung eines Wortes $w \in W$ in einem Vektorraum V wird dann repräsentiert durch eine Menge von Basisvektoren $\{\mathbf{e}_{(r,\,w')} \mid r \in R, w' \in W\}$. Solch ein Vektor dokumentiert die Zusammenhangsstärke zwischen einem Wort w und jedem Kontextwort w', welche in einer Relation r auftreten. Insbesondere wird ein Wort $w \in W$ mit einem Vektor $\mathbf{v}(w) \in V$ auf folgende Weise vereinigt:

$$\mathbf{v}(w) := \sum_{r \in R, w' \in W} f(w, r, w') * e_{(r,w')} \tag{3.1}$$

In der Formel 3.1 bezeichnet f eine Funktion, welche jedem Abhängigkeitstripel (w, r, w') ein Gewicht zuweist. Thater et al. benutzen hierfür den PMI-Wert (engl. *pointwise mutual information*), da der PMI (Church und Hanks, 1990) in ihren Experimenten gegenüber rohen Frequenzen für ein Tripel aus (w, r, w') bessere Ergebnisse liefert. Der PMI stellt ein Maß für den Zusammenhang zwischen zwei Variablen dar. Er gibt für beide Variablen jeweils an, wie viel Information das Auftreten einer Variable für das Auftreten der anderen bietet und ist folgendermaßen definiert (Reichel, 2008):

$$\boldsymbol{PMI}(w, r, w') = log \frac{p(w, w' \mid r)}{p(w, . \mid r) p(., w' \mid r)} \tag{3.2}$$

In Gleichung 3.2 stellen die Punkte relevante Variablen dar, über die iteriert werden muss.

Für gegebene Kontextwörter $w_1, ..., w_n$ und entsprechende syntaktische Relationen $r_1, ..., r_n$ wird ein kontextualisierter Vektor für w erhalten, indem für die Vektoren $\mathbf{v}_{r_i,w_i} (1 \leq i \leq n)$ eine Vektoraddition durchgeführt wird:

$$\mathbf{v}_{r_1,w_1,...,r_n,w_n}(w) := \sum_{i=1}^{n} \mathbf{v}_{r_i,w_i}(w) \tag{3.3}$$

Der resultierende Vektor $\mathbf{v}_{r_1,w_1,...,r_n,w_n}(w)$ ist dann die vollständig kontextualisierte Darstellung für das Wort **w**, welche Informationen zu allen Kontextwörtern beinhaltet.

3.3 Überwachte Methoden

Voraussetzung für überwachte Ansätze ist das Vorhandensein von manuell annotierten Korpora. Ein überwachtes Verfahren beinhaltet eine Trainingsphase, welcher sich eine Testphase anschließt. In der Trainingsphase wird ein mit Wortbedeutungen annotierter Korpus benötigt, von welchem syntaktische und semantische Feautures mittels maschinellem Lernen extrahiert werden, um einen Klassifizierer zu bilden. In der folgenden Testphase versucht der Klassifizierer anhand der das Zielwort umschließenden Kontextwörter ein angemessenes Substitut zu finden (Fulmari und Chandak, 2013).

3.3.1 Überwachtes Substitutionsmodell von Szarvas et al. (2013)

Einen überwachten Ansatz für lexikalische Substitution realisieren Szarvas et al. (Szarvas et al., 2013). Das System sieht davon ab, für jedes Wort einen separaten Klassifizierer zu benutzen und ist folglich auf jedes Wort im Vokabular anwendbar. Anstelle der Erlernung wortspezifischer Substitutionspattern wird ein globales Modell zur lexikalischen Ersetzung auf delexikalisierten (nicht lexikalen) Features trainiert. Das überwachte System kann Substitute generieren, welche nicht in den Trainingsdaten enthalten sind. Erreicht wird dies durch die Einbeziehung nicht lexikalischer Features aus verschiedenen Quellen. Das Modell benutzt die folgenden Featuregruppen:

- **Features aus lexikalischen Ressourcen:** WordNet als Quelle für Substitute und delexikalisierte Features
- **Korpus basierte Features:** Extraktion von Features für Wörter aus Abhängigkeiten
- **Lokale N-Gramm-basierte Features:** Nutzung von N-Gramm Informationen aus Web 1T
- **Flach syntaktische Features:** Nutzung von 1-3-Gramm Haupt-POS-Kategorien

Nach der Berechnung der verschiedenen Features wird mittels der Maximum-Entropie-Methode (MaxEnt) ein binärer Klassifizierer trainiert, welcher beurteilt, ob ein Substitut in einem bestimmten Kontext angemessen ist oder nicht. Die Ausgabe von MaxEnt ist eine Wahrscheinlichkeitsverteilung für jedes Paar aus Zielwort und Substitut, die für jede Instanz bei Betrachtung der Features, welche sowohl die Wörter, als auch ihre Kontexte beschreiben, angibt, wie gut sie als Substitut geeignet ist. Aus der Wahrscheinlichkeitsverteilung wird schließlich ein Ranking der Substitute generiert und lexikalische Substitution durchgeführt. Ihr System evaluieren sie auf dem LexSub Datensatz (siehe Kapitel 4.3) und erzielen bessere Ergebnisse gegenüber der Baseline. Bei der Baseline handelt es sich um Substitute, die aus WordNet 2.1 stammen. Tabelle 3.1 zeigt die erzielten Präzisionswerte dieses Ansatzes mittels *Generalized Average Precision* (GAP, siehe Kapitel 5.1) gegenüber der Baseline auf dem LexSub Datensatz.

	Substitute aus WordNet (GAP)	**Substitute aus LexSub Datensatz** (GAP)
Baseline	36.8%	46.9%
Szarvas et al.	43.8%	52.4%

Tabelle 3.1: Ergebnisse auf dem LexSub Datensatz

Einen Vergleich der Ergebnisse dieses Ansatzes zu früheren Arbeiten, unter anderem zu dem Vektorraummodel von Thater et al. (2011), auf dem LexSub Datensatz, bietet Tabelle 3.2. Die weiteren Arbeiten, die in Tabelle 3.2 aufgeführt sind, werden im Folgenden kurz erläutert.
Bei PadóErk10 handelt es sich um einen Exemplar basierten Ansatz für die Modellierung der Wortbedeutung im Kontext. Anstelle der Darstellung von Wort und Kontext als verschiedene Vektoren und späterer Kombination dieser, werden einige Wortvorkommen in ähnlichen Kontexten gesammelt und es werden lediglich diese Exemplare für die Repräsentation des Worts im Kontext herangezogen (Erk und Padó, 2010). Um einen Ansatz basierend auf einem Latente-Variablen-Modell handelt es sich bei DinuLapata. Die Wortbedeutung wird hierbei mittels einer Wahrscheinlichkeitsverteilung über verschiedene latente Variablen dargestellt. Für eine kontextualisierte Darstellung der Wortbedeutung wird das Modell auf den das Zielwort umgebenden Kontextwörtern konditioniert (Dinu und Lapata, 2010).

	Substitute aus LexSub Datensatz (GAP)
PadóErk10	38.6%
DinuLapata	42.9%
Thater11	51.7%
Baseline	46.9%
Szarvas et al.	52.4%

Tabelle 3.2: Vergleich zu früheren Ansätzen auf dem LexSub Datensatz

4 Datensätze und Ressourcen

Die Verfügbarkeit von reichhaltigem lexikalischen Wissen ist grundlegend für viele Anwendungen in der maschinellen Sprachverarbeitung. Solche Anwendungen sind unter anderem die Textzusammenfassung (Nastase, 2008), Disambiguierung von Named Entities (Faruqui et al., 2010), (Bunescu und Pasca, 2006), Beantwortung von Fragen (Harabagiu et al., 2000), (Lita et al., 2004), Textkategorisierung (Gabrilovich und Markovitch, 2007), (Navigli et al., 2011), (Wang und Domeniconi, 2008), Auflösung von Koreferenzen (Ponzetto und Strube, 2007), sowie die Plagiatserkennung (Barrón-Cedeno et al., 2010). Dieses Wissen kann in verschiedenen Formen vorliegen. Hierzu gehören nebst weiteren unstrukturierte Textsammlungen, annotierte Korpora, die in der Computerlinguistik als Datensatz angesehen werden (Napoles et al., 2012), Thesauri (Roget, 2008), maschinenlesbare Wörterbücher (Paul et al., 1978), sowie reichhaltige lexikalische Datenbanken. In der Fachliteratur herrscht Einigkeit darüber, dass die Existenz von umfangreichen Wissensquellen zu besseren Ergebnissen führt. Jedoch ist die Erstellung solcher lexikalischer Quellen äußerst zeit- und kostspielig.

Im Folgenden werden die in dieser Arbeit benutzten und erwähnten Datensätze und Ressourcen angeführt und erläutert. Insbesondere wird dabei auf ihre Zusammensetzung und Struktur eingegangen. Diese sind:

- **CoInCo - Concepts in Context**
- **LexSub Datensatz**
- **CatVar**
- **BabelNet**

4.1 CoInCo - Concepts in Context

Der CoInCo-Korpus[1] ist ein großer, alle Wörter umfassender Datensatz für die englische Sprache. Als ein Korpus für lexikalische Substitution umfasst er 35.000 Tokens laufenden Texts, entnommen von den newswire- und fiction-Teilen des frei zugänglichen MASC-Korpus[2]. Für die in diesem Korpus enthaltenen 15.629 Inhaltswörter (Nomen, Verben, Adjektive und Adverben) beinhaltet er mindestens 6 Substitute, welche durch Annotatoren in einem Crowdsourcing-Prozess[3] vorgeschlagen wurden. Bei

[1] Verfügbar auf: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/coinco.html
[2] http://www.anc.org/data/masc/
[3] Crowdsourcing bezeichnet den Prozess, bei dem Arbeit,hier die Annotation, auf eine große Zahl freiwilliger oder bezahlter Arbeiter verteilt beziehungsweise ausgelagert wird.

diesem Vorgang standen den Annotatoren neben dem Zielsatz zwei weitere Sätze als Diskurskontext zur Verfügung. Die Aufteilung der der 15.629 Inhaltswörter nach Wortarten listet Tabelle 4.1 auf.

Wortart	Anzahl in CoInCo
Nomen	7.117
Verb	4.617
Adjektiv	2.470
Adverb	1.425

Tabelle 4.1: Aufteilung von CoInCo nach Wortarten

In Abbildung 4.1 ist ein Ausschnitt von CoInCo illustriert. Dieser zeigt den Zielsatz (targetsentence) *"A mission to end a war "* mit dem Diskurskontext (postcontext). Daraufhin folgt eine Auflistung der Wörter mit IDs, sogenannte Tokens. Für jedes Token wird weiterhin seine Wortform, das zugehörige Lemma sowie die Wortart (engl. part of speech[4], kurz pos) angegeben. Komplettiert wird der Eintrag mit der Angabe der Substitute (substitutions). Weiterhin wird bei jedem Substitut das Lemma, die Wortart, sowie eine Frequenz (freq) angegeben, welche aussagt, wie viele Annotatoren dieses Substitut im Hinblick auf den Kontext als passend erachtet haben.

```
<document>
  <sent MASCfile="NYTnewswire9.txt" MASCsentID="s-r0" >
    <precontext>

    </precontext>
    <targetsentence>
    A mission to end a war
    </targetsentence>
    <postcontext>
    AUSTIN, Texas -- Tom Karnes was dialing for destiny, but not everyone wanted to cooperate.
    </postcontext>
    <tokens>
      <token id="XXX" wordform="A" lemma="a" posMASC="XXX" posTT="DT" />
      <token id="4" wordform="mission" lemma="mission" posMASC="NN" posTT="NN" problematic="no" >
        <substitutions>
          <subst lemma="calling" pos="NN" freq="1" />
          <subst lemma="campaign" pos="NN" freq="1" />
          <subst lemma="dedication" pos="NN" freq="1" />
          <subst lemma="devotion" pos="NN" freq="1" />
          <subst lemma="duty" pos="NN" freq="1" />
          <subst lemma="effort" pos="NN" freq="1" />
          <subst lemma="goal" pos="NN" freq="2" />
          <subst lemma="initiative" pos="NN" freq="1" />
          <subst lemma="intention" pos="NN" freq="1" />
          <subst lemma="movement" pos="NN" freq="1" />
          <subst lemma="plan" pos="NN" freq="2" />
          <subst lemma="pursuit" pos="NN" freq="1" />
          <subst lemma="quest" pos="NN" freq="1" />
          <subst lemma="step" pos="NN" freq="1" />
          <subst lemma="task" pos="NN" freq="2" />
        </substitutions>
      </token>
```

Abbildung 4.1: CoInCo - Ein Ausschnitt

[4]Bei der Wortartenannotierung (engl. Part-of-speech-Tagging) werden in einem natursprachlichen Text Wortarten zu jedem Wort zugeordnet. Die Zuordnung geschieht dabei auf Grundlage einer zuvor definierten Tagliste, welche Abkürzungen für die verschiedenen Wortarten beinhaltet.

4.2 LexSub Datensatz

Als Gold Standard wird in der Linguistik ein Datensatz bezeichnet, dessen Anreicherung mit zusätzlicher syntaktischer und semantischer Information (Annotation) von Experten für gut befunden wird und welcher sich folglich für Evaluationszwecke gut eignet. LexSub[5] ist ein solcher Datensatz und wurde von McCarthy und Navigli (McCarthy und Navigli, 2009) bei SemEval 2007[6] für die lexikalische Substitution eingeführt. Er beinhaltet 2010 Sätze für 201 Zielwörter aus verschiedenen Wortarten. Jeder Satz wurde dabei fünf Muttersprachlern vorgestellt, welche dann so viele Substitute beziehungsweise Paraphrasen wie möglich für das Zielwort im Kontext angegeben haben. Die Substitute enthalten zusätzlich ein Gewicht, welches der Zahl der Annotatoren entspricht, die dieses Substitut genannt haben.

```
<?xml version="1.0" ?>
<!DOCTYPE corpus SYSTEM "lexsub.dtd">

<corpus lang="english">
    <lexelt item="bright.a">
        <instance id="1">
            <context>During the siege , George Robertson had appointed Shuja-ul-Mulk ,
            who was a <head>bright</head> boy only 12 years old and the youngest
            surviving son of Aman-ul-Mulk , as the ruler of Chitral .</context>
        </instance>
        <instance id="2">
            <context>The actual field is not much different than that of a 40mm ,
            only it is smaller and quite a bit noticeably <head>brighter</head> ,
            which is probably the main benefit .</context>
        </instance>
        <instance id="3">
            <context>The roses have grown out of control , wild and carefree , their
            <head>bright</head> blooming faces turned to bathe in the early autumn sun .
            </context>
        </instance>
        <instance id="4">
            <context>He was <head>bright</head> and independent and proud .</context>
        </instance>
        <instance id="5">
            <context>In fact , during at least six distinct periods in Army history since
            World War I , lack of trust and confidence in senior leaders caused the so-called
            best and <head>brightest</head> to leave the Army in droves .</context>
        </instance>
        <instance id="6">
            <context>An evening of classical symphonic music , played by the next generation
            stars in the American orchestral scene , can be savored at the New World Symphony ,
            a special Miami institution that nurtures the best and <head>brightest</head> young
            symphonic musicians .</context>
        </instance>
```

Abbildung 4.2: Gold Standard - Kontextsätze

Abbildung 4.2 zeigt die ersten 6 der 10 Sätze für das Zielwort *bright*, welches hervorgehoben ist. Der Kontext für die Instanz-ID 3 ist dabei der Satz *"He was bright and independent and proud"*. Tabelle 4.2 listet die genannten Substitute für das Zielwort *bright* mit den Gewichten für die ersten 6 Sätze.

[5]Verfügbar auf: http://nlp.cs.swarthmore.edu/semeval/tasks/task10/data.shtml
[6]Bei SemEval, ehemals SENSEVAL, handelt es sich um eine Serie von fortlaufenden Workshops, welche sich mit der Evaluation von Systemen auseinandersetzen, die sich mit der Interpretation natürlicher Sprache befassen.

SATZ-ID	ZIELWORT	WORTART	SUBSTITUTE MIT (GEWICHT)
1	bright	Adjektiv	intelligent(3), clever(3), smart(1)
2	bright	Adjektiv	luminous(2), well-lit(1), clear(1), light(1)
3	bright	Adjektiv	colourful(2), brilliant(1), gleam(1), luminous(1)
4	bright	Adjektiv	intelligent(3), clever(3)
5	bright	Adjektiv	intelligent(3), clever(2), most able(1), capable(1)...
6	bright	Adjektiv	talented(3), up-and-coming(1), gifted(1), ...

Tabelle 4.2: Gold Standard: Substitute mit Gewicht

4.3 CatVar

CatVar ist eine umfangreiche Datenbank, welche kategorische Variationen für englische Lexeme beinhaltet (Habash und Dorr, 2003). Eine kategorische Variation eines Wortes mit einer bestimmten Wortart, ist ein Wort, welches in einer derivativen Beziehung mit dem ursprünglichen Wort steht. Die Variation kann dabei eventuell einer anderen Wortart angehören. So sind zum Beispiel die Wörter *Hunger* (Nomen), *hungern* (Verb) und *hungrig* (Adjektiv) kategorische Variationen von einander. Die CatVar-Datenbank wurde mithilfe einiger lexikalischer Quellen und Algorithmen entwickelt und beinhaltet in ihrer zweiten Version 62.232 Cluster mit 96.368 spezifischen Lexemen. Tabelle 4.3 zeigt die pronzentualen Anteile der verschieden Wortarten in der Datenbank.

Wortart	Pronzentueller Anteil
Nomen	62%
Adjektiv	24%
Verb	10%
Adverb	4%

Tabelle 4.3: Zusammensetzung von CatVar nach Wortarten

4.4 BabelNet

BabelNet[7] ist ein enzyklopädisches Wörterbuch (Navigli, 2013). Als ein sehr großes multilinguales semantisches Netzwerk verfügt es über eine große Überdeckung und beinhaltet unter anderem lexikographisches und enzyklopädisches Wissen aus WordNet und Wikipedia. BabelNet enthält ferner lexikalische Information zu vielen Sprachen, welche mittels maschineller Übersetzung gewonnen wurde. Es beinhaltet mehr als 13 Millionen Einträge, die als *Babel Synsets* bezeichnet werden. Jedes dieser *Babel Synsets* stellt eine Wortbedeutung sowie die zu dieser Bedeutung gehörenden Synonyme

[7] Verfügbar auf: http://babelnet.org/

dar. In den folgenden Unterkapiteln wird auf die zwei Hauptquellen (WordNet und Wikipedia), aus denen BabelNet entstanden ist, eingegangen. Die weiteren Quellen sind:

- **Open Multilingual WordNet**[8]**:** Eine Sammlung von Wortnetzen in anderen Sprachen
- **OmegaWiki**[9]**:** Ein großes, kollaboratives und vielsprachiges Wörterbuch
- **Wiktionary**[10]**:** Ein kollaboratives Projekt zur Erstellung eines multilingualen Wörterbuchs
- **Wikidata**[11]**:** Eine freie maschinenlesbare Wissensbasis

4.4.1 WordNet

WordNet ist die bekannteste lexikalische Datenbank für die englische Sprache und beinhaltet Einträge zu den vier Wortarten Nomen, Verben, Adjektive und Adverben. Aufgrund ihrer freien Verfügbarkeit kommt diese Quelle häufig in Anwendungen im Bereich der maschinellen Sprachverarbeitung zum Einsatz. WordNet repräsentiert einen Begriff (Konzept) als ein Synonymset (Synset genannt). Ein Synset besteht aus Begriffen, welche miteinander die gleiche Bedeutung teilen. Zu jedem Synset bietet WordNet eine kurze textuelle Definition beziehungsweise Glosse sowie Beispielsätze, in denen dieser Begriff vorkommt.

- S: (n) **bank#1** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution#1, **bank#2**, banking concern#1, banking company#1 (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank#3** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank#4** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank#5** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank#6** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank#7**, cant#2, camber#2 (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank#2, coin bank#1, money box#1, **bank#8** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank#9**, bank building#1 (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank#10** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Abbildung 4.3: WordNet - Synsets zu *bank*

[8] http://compling.hss.ntu.edu.sg/omw/
[9] http://www.omegawiki.org/
[10] https://de.wiktionary.org/
[11] https://www.wikidata.org/

Abbildung 4.3 zeigt die Ausgabe der Websuche von WordNet 3.1 zu dem Begriff *bank*. Diese listet alle Synsets auf. Ein Synset wird mit einem blauen *S* am Anfang der Zeile angeführt. Diesem folgt ein kleiner roter Buchstabe, welcher die Wortart angibt (im Beispiel jeweils ein *n* für noun = Nomen). Der Abbildung kann man entnehmen, dass das Wort *bank* zehn verschiedene Bedeutungen annehmen kann. Das zweite Synset (**bank#2**) enthält die Synonyme *depository financial institution#1*, *bank#2*, *banking concern#1* und *banking company#1*. Jedem Synonym folgt eine Nummer, die angibt, welche Bedeutungsvariente des Wortes in einem Synset gemeint ist. Die textuelle Definition für dieses Synset lautet: "*a financial institution that accepts deposits and channels the money into lending activities*", sowie zwei Beispielssätze: "*he cashed a check at the bank; that bank holds the mortgage on my home*". Eine weitere Eigenschaft von WordNet ist das Vorhandensein von lexikalischen und semantischen Relationen zwischen Synsets. Hierzu gehören unter anderem die *is-a* Beziehung, welche Hyperonymie (Generalisierung) oder Hyponymie (Spezialisierung) angibt, die *instance-of* Beziehung, welche angibt, zu welcher Klasse eine Named Entity gehört und die *part-of* Beziehung, welche Meronymie und Holonymie charakterisiert. Die Tabelle 4.4[12] zeigt die Zusammensetzung von WordNet 3.0. Sie zeigt die Anzahl der in dieser Ressource vorhandenen Wörter nach Wortarten, die Anzahl der jeweiligen Synsets, die Wortbedeutungspaare, sowie den durchschnittlichen Grad an Polysemie.

Wortart	# Wörter	# Synsets	# Wortsinn Paare	Durschnittl. Polysemie
Nomen	117.798	82.115	146.312	1.24
Verb	11.529	13.767	25.047	2.17
Adjektiv	21.479	18.156	30.002	1.40
Adverb	4.481	3.621	5.580	1.25
Summe	**155.287**	**117.659**	**206.941**	**1.52**

Tabelle 4.4: WordNet 3.0: Zusammensetzung der Datenbank

4.4.2 Wikipedia

Wikipedia[13] ist eine vielsprachige und frei zugängliche web-basierte Enzyklopädie. Sie ist ein von Freiwilligen gemeinschaftlich erzeugtes Medium und bietet eine umfassende Sammlung an enzyklopädischem Wissen. Die Einträge in Wikipedia sind in Form von Artikeln organisiert und jeder dieser Artikel entspricht einer Wikiseite (Navigli und Ponzetto, 2012). Ein Artikel bietet Information zu einem Begriff (vgl. *bank*) oder einer Named Entity (vgl. *Bank of England*). Der Titel einer Wikiseite besteht aus dem Lemma des Begriffs, zu dem Information angegeben wird, und eventuell aus einer Erweiterung in Klammern, falls es sich bei dem Begriff um ein ambiges Wort handelt. Wikiseiten sind untereinander mittels verschiedener Relationen verlinkt. Hierzu gehören:

- **Weiterleitungen:** Diese leiten einen Seitentitel auf eine andere Seite um. Oft handelt es sich dabei um Synoyme.

[12] Quelle: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
[13] https://en.wikipedia.org/

- **Begriffsklärungsseiten:** Falls es sich bei einem Stichwort um ein Polysem oder Homonym handelt, bieten diese Seiten Links zu den verschiedenen Bedeutungen.
- **Interne Links:** Wikiseiten enthalten sogenannte "Wikilinks", welche auf ähnliche Konzepte oder mehr Information zu einem Begriff verweisen.
- **Links zu anderen Sprachen:** Viele Artikel enthalten Links zu dem Artikel in anderen Sprachen.
- **Kategorien:** Diese sind ein Mittel, um Wikiseiten nach bestimmten Merkmalen einzuordnen.

BabelNet repräsentiert Wissen in Form eines gelabelten gerichteten Graphen *G = (V, E)*, wobei V die Menge der Knoten und $E \subseteq V \times R \times V$ die Menge der Kanten, die zwei Bedeutungen verbindet, darstellt. Ein Knoten ist zum Beispiel das Wort *play* beziehungsweise die Named Entity *Shakespeare*. Jede Kante verfügt über eine semantische Relation aus R der Form *is-a, part-of, ...,* ε, wobei ε eine unspezifizierte Relation darstellt (Navigli und Ponzetto, 2012). Die Konzepte und Relationen in BabelNet stammen aus WordNet und Wikipedia, insbesondere aus WordNet alle Wortbedeutungen als Konzepte und lexikalische und semantische Verweise zwischen Synsets als Relationen. Des Weiteren beinhaltet BabelNet enzyklopädische Einträge aus Wikipedia, wobei die Wikiseiten Konzepten und Wikilinks Relationen entsprechen. BabelNet wurde erstellt, indem Wikipedia auf WordNet gemappt wurde. Um dies zu erreichen wurde jede Wikiseite einem Begriff in WordNet zugeordnet. Formal betrachtet sieht das Mapping wie folgt aus:

$$\mu : Bedeutungen_{Wiki} \rightarrow Bedeutungen_{WN} \cup \{\varepsilon\}, \tag{4.1}$$

so dass für jede Wikiseite $w \in Bedeutungen_{Wiki}$, gilt:

$$\mu(w) = \begin{cases} s \in Bedeutungen_{WN}(w) & \text{falls eine Zuordnung möglich} \\ \varepsilon & \text{sonst,} \end{cases} \tag{4.2}$$

wobei $Bedeutungen_{WN}(w)$ die verschiedenen Bedeutung des Lemmas von w in WordNet darstellt. Tabelle 4.5 zeigt die Überdeckung von Babelnet 2.5.1[14] für einen Auszug der Sprachen.

Sprache	Lemmas	Synsets	Wortbedeutungen	Durchschnittliche Anzahl an Synonymen pro Babelsynset
Englisch	8368007	4107164	11029757	2.69
Französisch	2973918	1889526	4203324	2.22
Spanisch	2884159	1710523	3908085	2.28
Deutsch	2866334	1913433	3894287	2.04

Tabelle 4.5: BabelNet 2.5.1: Einige Sprachen und ihre Abdeckung

[14]Quelle: http://babelnet.org/2.5.1/stats

5 Eigener Ansatz

Die Zielsetzung dieser Arbeit liegt in der Durchführung der lexikalischen Substitution auf dem CoInCo-Korpus (siehe Kapitel 4.4). Lexikalische Substitution besteht aus der Generierung und anschließendem Ranking von Substitutionskandidaten für jede Instanz eines Zielworts im Hinblick auf ihre Angemessenheit in dem jeweiligen Kontext. Für das Ranking wird der vereinfachte Lesk-Algorithmus verwendet. Mit diesem werden die gefundenen Synsets und somit auch die in diesen Synsets enthaltenen Synonyme für ein Target gerankt. Hierfür wird die Überlappung (Anzahl gemeinsamer Wörter) zwischen der textuellen Definition in einem Synset und den Kontextwörtern, die das Zielwort umgeben, gebildet. Alle Substitutionskandidaten in einem Synset bekommen folglich das Ranking des Synsets zugewiesen.
Der vereinfachte Lesk-Algorithmus als Pseudocode sieht wie folgt aus:

Vereinfachter Algorithmus von Lesk
Sei w eine Instanz eines Lemmas W Seien $s_1, ..., s_n$ eine Menge an Babel Synsets Sei K die Menge der Kontextwörter, die w umgeben **Für jedes** Babel Synset s_i von W **Berechne** *Überlappung(s_i, K)*, die Anzahl der gemeinsamen Wörter in den Glossen von s_i und den *Kontextwörtern* K um w herum **Bilde** geordnete Liste der Substitute in $s_1, ..., s_n$ anhand der *Überlappung(s_i, K)* **Ordne** Substitute in der geordneten Liste w zu

Zur Durchführung von Experimenten werden drei verschiedene Kontexte eines Zielworts (zu ersetzendes Wort, engl. *target*) definiert:

- **Kontext1:** Der Kontext besteht aus dem Satz, in dem das Target enthalten ist.
- **Kontext3:** Der Kontext besteht aus dem Satz, in dem das Target erscheint einschließlich dem Satz davor und danach im Dokument.
- **Kontext5:** Der Kontext besteht aus dem Satz, in dem das Target erscheint einschließlich zweier Sätze vor und nach diesem Satz.

5.1 Methodik

Für die Durchführung der lexikalischen Substitution wird für jedes Target aus CoInCo der Wortart Nomen, Adjektiv oder Adverb, die Menge seiner Babel Synsets und deren Glosse sowie Beispielsätze

aus BabelNet angefordert. Handelt es sich bei dem Target um ein Verb, so wird dieser mittels CatVar (siehe Kapitel 4.2) in ein entsprechendes Nomen konvertiert, sofern CatVar eins beinhaltet. Dieser Schritt wird vollzogen, um eine eventuell schlechte Abdeckung von Verben vorzubeugen. Liefert CatVar mehr als ein Nomen, so wird das Nomen mit der höchsten Frequenz in CoInCo gewählt und es wird entsprechend der anderen Wortarten weiter verfahren. Falls CatVar zu einem Target kein Nomen liefert, so wird mit dem Verb weiter verfahren. Für die Bildung der Überlappung mittels Lesk-Algorithmus müssen die textuellen Definitionen sowie Beispielsätze aus BabelNet in einem Zwischenschritt tokenisiert werden. Hierfür wird das Stanford CoreNLP Natural Language Toolkit verwendet[1], mit welchem neben einer Lemmatisierung[2] auch eine Wortartenannotierung vorgenommen wird (Manning et al., 2014). Für CoInCo ist dieser Zwischenschritt nicht notwendig, da diese Ressource aufgrund ihrer Struktur direkt verwendet werden kann.
Nachdem die Babel Synsets und die textuellen Definitionen, sowie die Beispielsätze bereitstehen, wird für jede der zuvor genannten Kontexte jeweils die Überlappung berechnet. Falls mehrere Glossen für ein Zielwort existieren, werden diese konkateniert. Die Anzahl der Wörter in der Überlappung wird als Grundlage für das Ranking der Substitute verwendet. Je größer die Schnittmenge aus Kontext und textueller Definition sowie den Beispielsätzen ist, desto höher steht das Substitut im Ranking. Ein Ranking wird auch für die annotierten Substitute aus CoInCo vorgenommen. Hierfür gilt als Grundlage die Frequenz, welche der Anzahl der Annotatoren entspricht, die ein Substitut in dem jeweiligen Kontext als passend erachtet haben. Für ein besseres Verständnis wird der verfolgte Ansatz im Folgenden anhand des Nomens *increase* mit der Token ID 4624[3] erläutert.

Der Zielsatz, Pre- und Postkontext zu *increase* sehen wie folgt aus:

- <targetsentence> Its cereal division realized higher operating profit on volume *increases*, but also spent more on promotion. </targetsentence>
- <precontext> For the **year**, pet food volume was flat, the company said. </precontext>
- <postcontext> The Continental Baking business benefited from higher margins on bread and on increased cake sales, it **added**. </postcontext>

Zu diesem Target werden zwei Synsets gefunden, welche eine Schnittmenge in den Wörtern von 2 beziehungsweise 1 haben. Zu dem Synset mit zwei Worten Überlappung gehören folgende Glossen:

- <gloss> A quantity that is **added**; "there was an addition to property taxes this **year**"; "they recorded the cattle's gain in weight over a period of weeks«/gloss>
- <gloss> An amount by which a quantity is increased. </gloss>
- <gloss> An amount by which a quantity is enlarged. </gloss>

Die zwei Wörter in der Schnittmenge sind **year** und **add**. Die zu diesem Synset gehörigen Substitutionskandidaten sind:

[1]Verfügbar auf: http://nlp.stanford.edu/software/corenlp.shtml
[2]Die Lemmatisierung beschreibt den Prozess, bei dem jedem Wort eines laufenden Textes seine Grundform (das Lemma) zugeordnet wird.
[3]Dieses Beispiel entstammt aus der Konfiguration aus 3 Kontextsätzen und ohne den Einsatz von CatVar.

- <sense> addition </sense>
- <sense> gain </sense>
- <sense> increase </sense>
- <sense> increment </sense>

Zu dem Synset mit einem Wort Überlappung gehören folgende Glossen:

- <gloss> A process of becoming larger or longer or **more** numerous or **more** important; "the increase in unemployment"; "the growth of population«/gloss>
- <gloss> The amount of increase. </gloss>
- <gloss> The increase of the quantity, of measurements. </gloss>

  Die Schnittmenge für dieses Synset besteht aus dem Wort **more**. Folgende Substitutionskandidaten liefert dieses Synset:

- <sense> growth </sense>
- <sense> increase </sense>
- <sense> increment </sense>

5.2 Einführung der Formel für die Evaluation

Nachdem die Substitute erarbeitet sind, wird für jedes Target die Präzision mittels *Generalized Avarage Precision* (kurz GAP) berechnet (Kishida, 2005), (Thater et al., 2011).

$$\textbf{GAP} = \frac{\sum_{i=1}^{n} I(x_i) p_i}{R'} \tag{5.1}$$

Hierbei entspricht x_i dem Gewicht des i'ten Elements im Gold Standard (hier CoInCo), falls es darin enthalten ist, Null andernfalls. In der Gleichung 5.1 ist $I(x_i)$ eine binäre Variable und ist wie folgt definiert:

$$I(x_i) = \begin{cases} 1 & \text{falls } x_i > 0 \\ 0 & \text{andernfalls} \end{cases} \tag{5.2}$$

Die Variable p_i gibt die Präzision an der Stelle i an:

$$p_i = \frac{\sum_{k=1}^{i} x_k}{i} \tag{5.3}$$

Und R' lässt sich folgendermaßen berechnen:

$$R' = \sum_{i=1}^{R} I(y_i)\bar{y}_i \tag{5.4}$$

Hierbei entspricht $\bar{y}_i$ dem Durchschnittsgewicht (CoInCo-Frequenzen) der ideal gerankten Liste $y_1, ..., y_i$. Mittels Einsetzen lässt sich die GAP-Formel auch folgendermaßen schreiben:

$$\textbf{GAP} = \frac{\sum_{i=1}^{n} I(x_i)(\frac{\sum_{k=1}^{i} x_k}{i})}{\sum_{i=1}^{R} I(y_i)\bar{y}_i} \tag{5.5}$$

Die Anwendung dieser Formel erfordert ein eindeutiges Ranking der erarbeiteten Substitute. Da die Substitute in diesem Ansatz jedoch aufgrund von Schnittmengen gebildet werden und viele Synsets mehrere Substitutskandidaten enthalten, entstehen Substitute mit identischem Ranking. Um dieses Problem zu lösen, wird die Formel 5.3 zur Berechnung der Präzision an Stelle i leicht abgeändert. Das i, welches die Position des Substitutskanditaten in der erarbeiteten Liste angibt wird durch i' ersetzt. Die Berechnung von i' wird im Folgenden anhand eines Beispiels erklärt.

Betracht man das Nomen *increase* mit der ID 4624 aus CoInCo, so liefert der Algorithmus 6 Synsets, von denen 4 nicht betrachtet werden, weil ihre Schnittmengen leer sind. Die anderen 2 Synsets haben dabei einen Schnitt aus zwei beziehungsweise einem Wort und folglich die Wertigkeit zwei respektive eins. In dem Synset mit der Wertigkeit 2 befinden sich die Substitute *addition, gain, increase* und *increment*, in dem zweiten Synset die Kandidaten *growth, increase* und *increment*. In beiden genannten Sysnsets wird das Wort *increase* nicht betrachtet, da es sich um das zu ersetzende Zielwort handelt. Ferner wird *increment* aus dem zweiten Synset gestrichen, da es sich bereits im höher gewichteten ersten Synset befindet. Die Kandidaten im ersten Synset bekommen als i die Werte 1, 2 und 3 und als i' $\frac{(1+2+3)}{3} = 2$ und *growth* im zweiten Synset als i' $\frac{4}{1} = 4$. Tabelle 5.1 verdeutlicht die Berechnung von i'.

Substitut (CoInCo)	**Gewicht**	**Substitut (eigene Erarbeitung)**	**Gewicht**	i'	**Berechnung von** i'
gain	4	addition	2	2	(1+2+3)/ 3
addition	3	gain	2	2	(1+2+3)/ 3
rise	2	increment	2	2	(1+2+3)/ 3
growth	1	growth	1	4	4 / 1
improvement	1				
increment	1				

Tabelle 5.1: Berechnung von i'

Somit sieht die benutzte GAP-Formel wie folgt aus:

$$\textbf{GAP} = \frac{\sum_{i=1}^{n} I(x_i)(\frac{\sum_{k=1}^{i} x_k}{i'})}{\sum_{i=1}^{R} I(y_i)\bar{y}_i} \tag{5.6}$$

Die Berechnung von GAP für das Target *increase* sieht wie folgt aus:

$$\textbf{GAP}(\textit{increase}) = \frac{\frac{3}{2} + \frac{7}{2} + \frac{8}{2} + \frac{9}{4}}{4 + \frac{7}{2} + \frac{9}{3} + \frac{10}{4} + \frac{11}{5} + \frac{12}{6}} = \frac{225}{344} = 0.6540697 = 65.40697\%$$

Nach der Berechnung der GAP-Werte für jedes Zielwort aus CoInCo wird die Präzision des verfolgten Ansatzes ermittelt, in dem die Summe aller GAP-Werte durch die Anzahl der Zielwörter geteilt wird.

5.3 Technische Umsetzung

Die Umsetzung des Ansatzes erfolgte mit der Programmiersprache Java. Das Programm verfügt über eine graphische Benutzeroberfläche zur Analyse und erzeugt ferner eine XML-Datei, welche eine angereicherte Version von CoInCo darstellt. Die Erweiterung enthält zu jedem Token die gefundenen Synsets, welche aus Glossen, Beispielsätzen und Substitutionskandidaten bestehen, sowie die erzielten GAP-Werte. Die Struktur der erzeugten XML-Datei zeigt Abbildung 5.1.

```
<document tokenCount="" totalratio="" systemsGAP="">
    <postcontext>          </postcontext>
    <precontext>           </precontext>
    <targetsentence>       </targetsentence>
    <tokens>
<token id="" lemma="" posMASC="" posTT="" problematic="" wordform="" gap="">
        <substitutions>
          <subs freq="" lemma=""/>
        </substitutions>
        <synsets>
          <synset intersection="" intersectionset="">
            <glosses>
<gloss>      </gloss>
            </glosses>
            <senses>
<sense>      </sense>
            </senses>
          </synset>
```

Abbildung 5.1: Struktur der erzeugten XML-Datei

Die graphische Oberfläche ist in Abbildung 5.2 dargestellt. Hier kann der Benutzer unter *File* auswählen, ob CatVar bei der Analyse verwendet werden soll. Anschließend kann er unter *Options* auswählen, wie viele Sätze als Kontext betrachtet werden sollen. Die Analyse startet, sobald der Benutzer zu der Datei *coinco.xml* navigiert und diese öffnet.

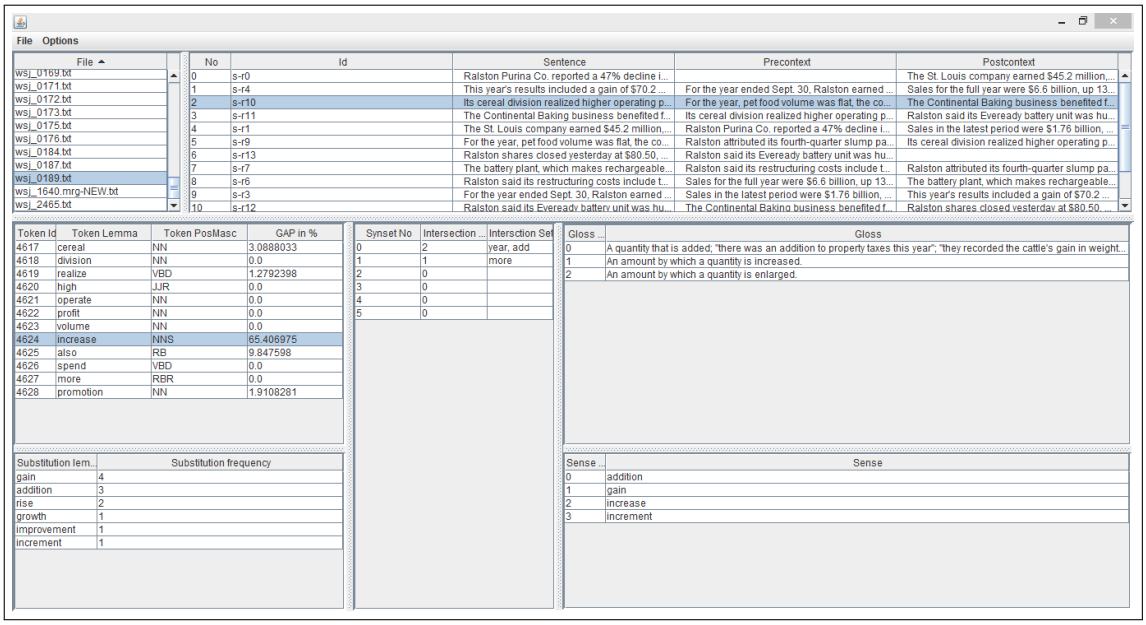

Abbildung 5.2: Graphische Benutzeroberfläche

6 Evaluation

Die Evaluation des verfolgten Ansatzes erfolgt mittels der GAP-Formel (*Generalized Average Precision*) gegen die im CoInCo-Korpus annotierten Substitute. Mittels GAP lässt sich die Präzision von erarbeiteten Substituten für ein Zielwort bestimmen. Hierbei erstrecken sich diese Werte zwischen 0 und 1, wobei der Wert 1 eine hundertprozentige Präzision bedeutet. Solch ein Wert wird erreicht, wenn alle richtigen Substitute die falschen und alle höher gewichteten die niedriger gewichteten überragen. Im vorangegangenen Kapitel wurde bereits der GAP-Wert für *increase* berechnet. Dieses Kapitel gibt nun die erzielten Ergebnisse für die verschiedenen Konfigurationen des gesamten Systems an und diskutiert sie.

Die folgende Tabelle gibt die Ergebnisse dieses Ansatzes an.

	Kontext1	**Kontext3**	**Kontext5**
Mit CatVar	0.752%	1.247%	1.520%
Ohne CatVar	1.046%	1.717%	**2.083%**

Tabelle 6.1: Ergebnisse für verschiedene Konfigurationen

In der besten Konfiguration wird eine Präzision von lediglich **2.083%** erzielt. Dieser Wert ist unter anderem deshalb so niedrig, weil der Algorithmus nur für 3635 Zielwörter Substitute generiert, was einem Anteil von 23.4% entspricht. Im Umkehrschluss bedeutet das, dass der Algorithmus für 11994 (76,6%) der Zielwörter keine Substitute generieren kann. Die Hauptursache hierfür liegt in der Kürze beziehungsweise Knappheit der textuellen Definitionen in den Synsets. In den meisten Fällen sind die Schnittmengen leer, was dazu führt, dass keine Substitute erzeugt werden. Ferner kommt es auch vor, dass Synsets als Synonym lediglich das Target enthalten, welches natürlich auch nicht als Substitut in Frage kommt.

Betrachtet man nur die Targets, für die der Algorithmus Substitute findet, so liegt die Präzision bei **8.955%**. Die Ergebnisse aus Tabelle (6.1) zeigen auch, dass die Verwendung von CatVar sich negativ auf die Präzision auswirkt. Die höchste Präzision mit CatVar wird unter Betrachtung von 5 Kontextsätzen erreicht. In dieser Konfiguration wird eine Präzision von 1.520% erreicht. Die Präzision steigt hier minimal, wenn man die 19 Zielwörter nicht betrachtet, die mittels CatVar in Nomen umgewandelt werden, für die jedoch keiner der Substitute aus BabelNet in ein Verb zurück gewandelt werden kann. Ferner werden für 62 Verben keine Nomen gefunden und es wird mit dem Verb weiter verfahren. Das heißt, in dieser Konfiguration kommt CatVar für 81 von 4.617 Verben aus CoInCo nicht zum Einsatz. Die Präzision beträgt dann 1.522%.
Um die Beiträge der einzelnen Quellen aus BabelNet zu untersuchen wurden weitere Experimente

mit Einschränkung in den verwendeten Quellen aus BabelNet durchgeführt. Dies wurde für die Konfiguration aus 5 Kontextsätzen und ohne die Verwendung von CatVar durchgeführt. Diese Ergebnisse veranschaulicht Tabelle 6.2.

Ohne Verwendung von	Generalized Average Precision
WordNet	1.779%
OmegaWiki	1.983%
Open Multilingual WordNet	2.083%
Wiktionary	2.083%
Wikipedia Seiten	2.087%
Wikipedia Weiterleitungen	2.096%
WikiData	2.085%

Tabelle 6.2: Ergebnisse ohne Verwendung einzelner Quellen

Aus Tabelle 6.2 kann man entnehmen, dass WordNet gefolgt von OmegaWiki die größten Beiträge liefert. Die anderen Quellen tragen nichts zum erzielten Ergebnis bei und wirken sich sogar minimal negativ auf die Ergebnisse aus. Den größten negativen Effekt liefern die Wikipedia Weiterleitungen. Ohne die Verwendung dieser verbessert sich die GAP von 2.083% minimal auf 2.096%. Dies hängt damit zusammen, dass aus diesen Quellen zusätzliche falsche Substitute generiert werden, welche sich negativ auf die Präzision auswirken.

In einem weiteren Experiment wurde die lexikalische Substitution lediglich mit einer Quelle aus BabelNet durchgeführt. Die Ergebnisse dieses Experiments sehen wie folgt aus:

Nur Verwendung von	Generalized Average Precision
WordNet	1.984%
OmegaWiki	0.704%
Open Multilingual WordNet	0.0%
Wiktionary	1.415%
Wikipedia Seiten	0.175%
Wikipedia Weiterleitungen	0.117%
WikiData	0.161%

Tabelle 6.3: Ergebnisse mit Verwendung einzelner Quellen

Betrachtet man die Ergebnisse in der Tabelle 6.3, so fällt auf, dass die alleinige Nutzung von WordNet nahezu die gleiche Präzision erreicht wie die höchste Präzision aus Tabelle 6.1. Dies hängt mit der besseren Abdeckung dieser Ressource gegenüber den anderen zusammen. Eine weitere Erkenntnis ist, das Open Multilingual WordNet keine Ergebnisse liefert und folglich überflüssig ist. Des weiteren ist zu erkennen, dass die mit Wikipedia zusammenhängenden Ressourcen im Vergleich zu WordNet sehr schlechte Ergebnisse liefern.

In einem letzten Experiment wurde jeweils eine Menge verschiedener Quellen für die lexikalische Substitution verwendet. Die höchste *Generalized Average Precision* wurde hierbei durch die Verwendung der Quellen WordNet, OmegaWiki und Wiktionary erreicht und lag bei **2.113%**. Unter Vernachlässigung der Targets, für die keine Substitute generiert wurden, entspricht dieser Wert einer GAP von **9.180%**.

7 Zusammenfassung und Ausblick

Lexikalische Substitution (LS) ist ein relativ neues Paradigma zur Wortbedeutungsauflösung und unterscheidet sich von *Word Sense Disambiguation* dahingehend, dass LS die Quelle für mögliche Substitute nicht vorschreibt. Die Zielsetzung dieser Arbeit lag in der Durchführung einer vokabular globalen lexikalischen Substitution auf dem CoInCo-Korpus. Hierfür wurden neben CoInCo die Ressourcen BabelNet und CatVar eingesetzt. Mit Hilfe dieser wurden Substitute generiert. Das Ranking der erarbeiteten Substitute wurde mit dem vereinfachten Lesk-Algorithmus anhand der Schnittmengen zwischen den textuellen Definitionen und Beispielsätzen der Substitute einerseits und den Kontextwörtern um das zu ersetzende Wort aus CoInCo andererseits, durchgeführt. Ein wesentlicher Bestandteil dieser Arbeit lag in der Evaluation der erarbeiteten Substitute. Diese wurde mit der *Generalized Average Precision* gegen die im CoInCo annotierten Substitute durchgeführt.
Die erzielten Ergebnisse wurden angeführt und diskutiert. Ferner wurde auf ihr Zustandekommen eingegangen. Um die Beiträge der verschiedenen verwendeten Ressourcen zu ermitteln, wurden weitere Experimente durchgeführt, in denen die Ressourcen nicht benutzt (CatVar) beziehungsweise beschränkt wurden (BabelNet). Auch die Ergebnisse dieser Experimente wurden angegeben und diskutiert.
Zusammenfassend kann gesagt werden, dass die Zielsetzung erreicht wurde, jedoch die erzielten Ergebnisse nicht zufriedenstellend sind und weit unter den aktuellen in der Forschung erzielten Ergebnissen liegen. Dies liegt hauptsächlich an dem verwendeten Algorithmus. Mittels Lesk-Algorithmus lassen sich für die Mehrzahl der Targets keine Substitute generieren.

Ausblick

Um eine eventuelle Verbesserung der Ergebnisse zu erreichen, müssen für eine wesentlich größere Anzahl der Targets Substitute erzeugt werden. Dies kann mit dem Lesk-Algorithmus nur erreicht werden, wenn weitere nicht leere Schnittmengen gebildet werden können. Hierfür könnten unter anderem die Synonyme und Hypernyme der Targets und ihre textuellen Definitionen herangezogen werden, um die Wahrscheinlichkeit der Schnittmengenbildung zu erhöhen (Ponzetto und Navigli, 2010), (Hassan et al., 2007). Des weiteren könnten Synsets aus mehreren Ressourcen vereinigt werden, um Synsets zu vermeiden, die lediglich aus dem zu ersetzenden Wort bestehen.

Literaturverzeichnis

Agirre, E. und Edmonds, P. G. (2007). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer. (Zitiert auf den Seiten 17 und 19)

Auberle, A., Eickhoff, B., Knörr, E., Münzberg, F., Osterwinter, R., Rautmann, K., und Scholze-Stubenrecht, W. (2009). *Duden-die deutsche Rechtschreibung*. Dudenverlag. (Zitiert auf Seite 13)

Barrón-Cedeno, A., Rosso, P., Agirre, E., und Labaka, G. (2010). Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 37–45. Association for Computational Linguistics. (Zitiert auf Seite 23)

Bunescu, R. C. und Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16. (Zitiert auf Seite 23)

Bußmann, H. (2008). Lexikon der sprachwissenschaft. vierte, durchgesehene und bibliographisch ergänzte auflage unter mitarbeit von hartmut lauffer. (Zitiert auf Seite 9)

Carstensen, K.-U., Ebert, C., Endriss, C., Jekat, S., Klabunde, R., und Langer, H. (2010). Computerlinguistik und sprachtechnologie. *Eine Einführung*, 3. (Zitiert auf den Seiten 10 und 15)

Church, K. W. und Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29. (Zitiert auf Seite 20)

Dinu, G. und Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. Association for Computational Linguistics. (Zitiert auf Seite 22)

Erk, K. und Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the acl 2010 conference short papers*, pages 92–97. Association for Computational Linguistics. (Zitiert auf Seite 22)

Faruqui, M., Padó, S., und Sprachverarbeitung, M. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS*, pages 129–133. (Zitiert auf Seite 23)

Fellbaum, C. (1998). *WordNet*. Wiley Online Library. (Zitiert auf Seite 10)

Fulmari, A. und Chandak, M. B. (2013). A survey on supervised learning for word sense disambiguation. *International Journal of Advanced Research in Computer & Communication Engineering*, 2(12):4667–4670. (Zitiert auf Seite 21)

Gabrilovich, E. und Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611. (Zitiert auf Seite 23)

Habash, N. und Dorr, B. (2003). Catvar: A database of categorial variations for english. In *Proceedings of the MT Summit*, pages 471–474. (Zitiert auf Seite 26)

Harabagiu, S. M., Moldovan, D. I., Paşca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, C. R., Rus, V., und Morărescu, P. (2000). Falcon: Boosting knowledge for answer engines. (Zitiert auf Seite 23)

Hassan, S., Csomai, A., Banea, C., Sinha, R., und Mihalcea, R. (2007). Unt: Subfinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 410–413. Association for Computational Linguistics. (Zitiert auf den Seiten 18 und 41)

Hornby, A. S. und Wehmeier, S. (1995). *Oxford advanced learner's dictionary*, volume 1430. Oxford University Press Oxford. (Zitiert auf Seite 18)

Kishida, K. (2005). *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan. (Zitiert auf Seite 33)

Kremer, G., Erk, K., Padó, S., und Thater, S. (2014). What substitutes tell us-analysis of an all-words lexical substitution corpus. In *Proceedings of EACL*. (Zitiert auf Seite 9)

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM. (Zitiert auf Seite 17)

Lewandowski, T. (1994). Linguistisches wörterbuch. in 3 bdn., 6. *Aufl.. Bd*, 3. (Zitiert auf Seite 9)

Lita, L. V., Hunt, W. A., und Nyberg, E. (2004). Resource analysis for question answering. In *Proceedings of the acl 2004 on interactive poster and demonstration sessions*, page 18. Association for Computational Linguistics. (Zitiert auf Seite 23)

Löbner, S. (2003). *Semantik: eine Einführung*. Walter de Gruyter. (Zitiert auf den Seiten 13, 14 und 15)

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., und McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. (Zitiert auf Seite 32)

McCarthy, D. (2002). Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 109–115. Association for Computational Linguistics. (Zitiert auf Seite 11)

McCarthy, D. und Navigli, R. (2009). The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159. (Zitiert auf den Seiten 10, 11 und 25)

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., und Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244. (Zitiert auf Seite 10)

Napoles, C., Gormley, M., und Van Durme, B. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics. (Zitiert auf Seite 23)

Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 763–772. Association for Computational Linguistics. (Zitiert auf Seite 23)

Navigli, R. (2013). A quick tour of babelnet 1.1. In *Computational Linguistics and Intelligent Text Processing*, pages 25–37. Springer. (Zitiert auf Seite 26)

Navigli, R., Faralli, S., Soroa, A., de Lacalle, O., und Agirre, E. (2011). Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2317–2320. ACM. (Zitiert auf Seite 23)

Navigli, R. und Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250. (Zitiert auf den Seiten 28 und 29)

Ng, H. T. und Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47. Association for Computational Linguistics. (Zitiert auf Seite 10)

Paul, P. et al. (1978). Longman dictionary of contemporary english. *England: Longman Group Limited*. (Zitiert auf Seite 23)

Ponzetto, S. P. und Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1522–1531. Association for Computational Linguistics. (Zitiert auf Seite 41)

Ponzetto, S. P. und Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212. (Zitiert auf Seite 23)

Reichel, U. (2008). Statistische sprachmodelle. (Zitiert auf Seite 20)

Roget, P. M. (1911). *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company. (Zitiert auf Seite 18)

Roget, P. M. (2008). *Roget'S International Thesaurus, 3/E***. Oxford and IBH Publishing. (Zitiert auf Seite 23)

Szarvas, G., Biemann, C., Gurevych, I., et al. (2013). Supervised all-words lexical substitution using delexicalized features. In *HLT-NAACL*, pages 1131–1141. (Zitiert auf den Seiten 11, 17 und 21)

Thater, S., Fürstenau, H., und Pinkal, M. (2011). Word meaning in context: A simple and effective vector model. In *IJCNLP*, pages 1134–1143. (Zitiert auf den Seiten 19 und 33)

Wang, P. und Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721. ACM. (Zitiert auf Seite 23)

Weaver, W. (1949). Translation. *Mimeographed, 12 pp. Reprinted in Willam N. Locke & Donald A. Booth, eds. 1955. Machine Translation of Languages, 15-23. New York: John Wiley & Sons.* (Zitiert auf Seite 10)

Zhou, X. und Han, H. (2005). Survey of word sense disambiguation approaches. In *FLAIRS Conference*, pages 307–313. (Zitiert auf Seite 19)

Alle URLs wurden zuletzt am 15. 09. 2015 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift