

Tim vor der Brück

Wissensakquisition mithilfe maschineller Lernverfahren auf tiefen semantischen Repräsentationen

RESEARCH



Springer Vieweg

Wissensakquisition mithilfe maschineller Lernverfahren auf tiefen semantischen Repräsentationen

Tim vor der Brück

Wissensakquisition mithilfe maschineller Lernverfahren auf tiefen semantischen Repräsentationen

Tim vor der Brück
Frankfurt am Main, Deutschland

Dissertation FernUniversität in Hagen, 2012

ISBN 978-3-8348-2502-5
DOI 10.1007/978-3-8348-2503-2

ISBN 978-3-8348-2503-2 (eBook)

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Springer Vieweg

© Springer Fachmedien Wiesbaden 2012

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Einbandentwurf: Künkellopka GmbH, Heidelberg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist eine Marke von Springer DE. Springer DE ist Teil der Fachverlagsgruppe Springer Science+Business Media
www.springer-vieweg.de

Danksagung

Ich danke Herrn Prof. Dr. Hermann Helbig für die Vergabe des interessanten Themas, für viele hilfreiche Diskussionen, Verbesserungsvorschläge sowie für die zahlreichen Vertragsverlängerungen, die diese Arbeit erst möglich gemacht haben. Zudem danke ich Herrn Dr. habil. Helmut Horacek für die Bereitschaft, als Zweitgutachter diese Arbeit zu begutachten sowie für viele hilfreiche Hinweise. Ferner danke ich den Mitarbeitern am Lehrstuhl *Intelligente Informations- und Kommunikationssysteme* für ihre Hilfsbereitschaft und Unterstützung und allen anderen, die mir bei meiner Arbeit geholfen haben, insbesondere (in alphabetischer Reihenfolge) Dr. Tiansi Dong, Dipl.-Ing. Christoph Doppelbauer, Christian Eichhorn, Dr. Ingo Glöckner, Dr. Sven Harttrumpf, Dipl.-Math. Tom Kollmar, Dr. Johannes Leveling, Dr. Rainer Osswald, Alexander Pilz-Lansley, M.A. Andy Lücking sowie meinen Eltern.

Inhaltsverzeichnis

1. Einleitung	1
1.1. Motivation	1
1.2. Arten von Wissens	2
1.3. Vorteile der Nutzung von Wissen bei der natürlichsprachlichen Verarbeitung	3
1.4. Vorteile einer automatischen Wissensakquisition	5
1.5. Einordnung der Arbeit	6
1.5.1. Thematische Einordnung der Arbeit	6
1.5.2. Organisatorische Einordnung der Arbeit	7
1.6. Zusammenfassung und Thesen	8
1.7. Aufbau der Arbeit	9
1.8. Typografische Konventionen	10
2. Typische Arten von Wissen	11
2.1. Synonymie	11
2.2. Subordination	11
2.3. Meronymie	13
2.4. Entailments und Paraphrasen	14
3. Grundlagen	17
3.1. MultiNet	17
3.2. Verwendete Ressourcen	30
3.2.1. Verwendete Korpora	30
3.2.2. Axiome	31
3.2.3. HaGenLex	31
3.3. Evaluation	33
3.3.1. Evaluationsverfahren	33
3.3.2. Evaluationsmaße	34

4. Stand der Forschung	37
4.1. Extraktion von Synonymen und Wortsynonymen	37
4.2. Extraktion von Wortsubordinationen	44
4.2.1. Extraktion von Wortsubordinationen durch Betrachtung syntagmatischer Relationen	45
4.2.2. Extraktion von Wortsubordinationen durch Betrachtung paradigmatischer Relationen	52
4.2.3. Extraktion von Wortsubordinationen durch Dokument- clustering	53
4.2.4. Logische Validierung	57
4.3. Extraktion von Meronymen und Wortmeronymen	58
4.4. Lernen von ontologischen Sorten und semantischen Merkmalen .	61
4.5. Lernen von Entailments und Bedeutungspostulaten	62
4.6. Weitere Arten der Wissensakquisition	68
4.7. Nachteile bisheriger Verfahren und Vorteile tiefer Verfahren . .	69
4.8. Verwendung und Anpassung von Verfahren aus dem Stand der Forschung	71
5. Relationsextraktion	73
5.1. Vorgehen	74
5.2. Extraktion von Subordinationsrelationen	75
5.2.1. Tiefe Extraktionsregeln	75
5.2.2. Flache Extraktionsregeln	88
5.2.3. Lernen der Extraktionsregeln	90
5.2.4. Vorteile und Nachteile tiefer Extraktionsregeln gegenü- ber flachen	108
5.2.5. Fehlerquellen bei der Anwendung der tiefen Extraktions- regeln	110
5.2.6. Filtern anhand ontologischer Sorten und semantischer Merkmale	116
5.2.7. Ablegen der Relation in der Datenbank	120
5.2.8. Validierung von Subordinationshypthesen	121
5.2.9. Validierungsmerkmale	137
5.2.10. Auswahl der Merkmale	142
5.2.11. Wahl der korrekten Unterrelation	143
5.2.12. Erkennung von Eigennamen	143

5.3.	Extraktion von Meronymrelationen	145
5.3.1.	Anwendung der Extraktionsregeln	146
5.3.2.	Filtern anhand ontologischer Sorten und semantischer Merkmale	152
5.3.3.	Validierungsmerkmale	155
5.3.4.	Wahl der korrekten Unterrelation	161
5.3.5.	Echte und in SUB eingebettete Meronymie	162
5.3.6.	Propagieren von Meronymen	164
5.4.	Extraktion von Synonymen	167
5.4.1.	Synonymextraktionsregeln	167
5.4.2.	Filtern anhand ontologischer Sorten und semantischer Merkmale	171
5.4.3.	Validierungsmerkmale	172
5.5.	Logische Ontologievalidierung mithilfe eines automatischen Be- weisers	175
5.6.	Extraktion von semantischen Relationen aus maschinenlesbaren Lexika	184
5.7.	Technischer Support für die Annotation der Daten	189
6.	Entailments	193
6.1.	Vorgehen	194
6.2.	Datenmodell	198
6.3.	Zusätzliches Verfahren zur Extraktion von Entailmenthypothesen	199
6.4.	Merkmale zur Konfidenzwertberechnung	204
6.4.1.	EM1: Vorkommen von generischen Begriffen	204
6.4.2.	EM2: Vorkommen von Namen	204
6.4.3.	EM3: Beschreibungslänge	205
6.4.4.	EM4: Iterationen	205
7.	Evaluation	207
7.1.	Evaluation der Extraktion von Hyponymen	207
7.1.1.	Evaluation des Lernens von Hyponymextraktionsregeln	207
7.1.2.	Anwendung der Hyponymextraktionsregeln und Validie- rung	210
7.2.	Evaluation der Extraktion von Meronymen	219
7.2.1.	Evaluation des Lernens von Meronymextraktionsregeln	219
7.2.2.	Anwendung von Meronymextraktionsregeln	219

7.3. Evaluation der Extraktion von Synonymen	227
7.4. Evaluation der logischen Ontologievalidierung	232
7.5. Evaluation der Extraktion von Entailments und Bedeutungspos- tulatn	234
7.6. Schlussfolgerungen aus der Evaluation	241
8. Zusammenfassung und Ausblick	243
8.1. Zusammenfassung der erreichten Leistungen	243
8.2. Gültigkeit der Thesen	243
8.3. Ausblick und weitere Forschung	247
A. Verwendete Extraktionsregeln zur Relationsextraktion	249
A.1. Tiefe Regeln zur Extraktion von Subordinationsbeziehungen . .	250
A.2. Flache Regeln zur Extraktion von Subordinationsbeziehungen .	252
A.3. Tiefe Regeln zur Extraktion von Meronymen	255
A.4. Flache Regeln zur Extraktion von Meronymen	257
B. Zur Relationsextraktion verwendete Axiome	259
C. Listen extrahierter Hypothesen (beste und schlechteste Hypothe- sen)	261
C.1. Beste Hyponymhypothesen	262
C.2. Schlechteste Hyponymhypothesen	264
C.3. Beste Meronymhypothesen	266
C.4. Schlechteste Meronymhypothesen	272
C.5. Beste Synonymhypothesen	274
C.6. Schlechteste Synonymhypothesen	276
D. Entailments-Ankerliste	279
E. Anwendungssysteme	283
E.1. Der Lesbarkeitsüberprüfer DeLite	283
E.2. SemDupl	291
E.3. LogAnswer	292
F. Glossar	295

Abbildungsverzeichnis

3.1. Relationshierarchie für SUB0, MERO und LEXCOMP	19
3.2. Hierarchie der ontologischen Sorten (Kopie aus [Hel08, Seite 399]) 20	
3.3. Semantisches Netz zu dem Satz “ <i>Der Lehrer fährt mit dem Auto nach München.</i> ”	26
3.4. Lexikoneintrag für den Begriff <i>bär.1.1</i>	32
4.1. Worthyponymextraktion unter Verwendung eines bilingualen Wörterbuches	51
4.2. Aufbau einer Taxonomie mithilfe von Textmining	54
4.3. Beispiel einer Begriffshierarchie aus Quan et al.	55
5.1. Architektur der Relationsextraktion von SemQuire	76
5.2. Beispiel eines semantischen Netzes zu dem Satz “ <i>Ein Wolkenkratzer bezeichnet ein sehr hohes Haus.</i> ”	78
5.3. Beispiel eines semantischen Netzes zu dem Satz “ <i>Der Minister und andere Politiker kritisierten das Gesetz.</i> ”	80
5.4. Tiefe Extraktionsregel zur Subordinationsextraktion (Prämisse ist gegeben als semantisches Netz)	81
5.5. Darstellung von Hyponymen und Instanzrelationen mit Namensangabe	84
5.6. Beispiel eines semantischen Netzes zu dem Satz “ <i>Barack Obama und andere Politiker kritisierten das Gesetz.</i> ”	85
5.7. Tokens für den Satz “ <i>Der Bundeskanzler und andere Politiker kritisierten das Gesetz.</i> ” wie vom WOCADI-Parser zurückgegeben 91	
5.8. Eine flache Extraktionsregel, die verwendet wird, um Subordinationshypothesen zu extrahieren.	92
5.9. Die Prämisse einer flachen Extraktionsregel wird mit der Tokenliste unifiziert.	93
5.10. Adjazenzmatrix des semantischen Netzes	94
5.11. Adjazenzmatrix des komprimierten semantischen Netzes	98

5.12. Kompression eines semantischen Netzes durch eine Extraktionsregelprämisse	99
5.13. Graph nach Entfernen der redundanten Kanten/Knoten	100
5.14. Beispiel für Strahlensuche	100
5.15. Berechnung des kürzesten Pfades gemäß dem Dijkstra-Algorithmus	105
5.16. Berechnung des kürzesten Pfades gemäß dem Dijkstra-Algorithmus (2)	105
5.17. Distanz der Kanten zum kürzesten Pfad	107
5.18. Anwendung der in Abbildung 5.19 dargestellten tiefen Extraktionsregel auf das semantische Netz zu dem Satz <i>“Er verkauft alle gebräuchlichen Streichinstrumente außer Celli.”</i>	111
5.19. Tiefe Extraktionsregel für die Hyponymextraktion, wobei die Prämisse als semantisches Netz gegeben ist	112
5.20. Semantisches Netz zu dem Satz: <i>“Sein Vater und andere Polizisten . . .”</i>	113
5.21. Semantisches Netz zu Satz <i>“Sie kauften Fleisch, Butter und andere Milchprodukte.”</i>	116
5.22. Vergleich einer Hyponymhypothese mit einem Begriff, der als Wortetikett ein Kompositum besitzt.	119
5.23. Widerspruchsbeweis Subordination	119
5.24. Verwendetes Datenmodell für das Ablegen von extrahierten Subordinationsrelationen in der Datenbank	122
5.25. Klassifizierung mithilfe einer Support-Vektor-Maschine. Eine Hyperebene teilt den Datensatz in zwei Bereiche ein.	123
5.26. Gemeinsamer Pfad in zwei semantischen Netzen der Länge vier. Die beiden semantischen Netze repräsentieren die Sätze (links) <i>“Der Minister und andere Politiker kritisierten das Gesetz”</i> sowie (rechts) <i>“Er kaufte ein Cello und andere Instrumente.”</i>	127
5.27. Pseudocode für die Berechnung der gemeinsamen Wege	131
5.28. Anwendung der tiefen Extraktionsregel DM_4 auf das semantische Netz zu dem Satz <i>“Lindenthal ist ein Stadtteil von Köln”</i>	148
5.29. Semantisches Netz zu dem Satz <i>“Apfelsaftschorle ist eine Mischung aus Mineralwasser und Apfelsaft.”</i>	151
5.30. Grafische Illustration des Beweises der Propagierung von Meronymen	165
5.31. Propagieren von Meronymen	166

5.32. Semantisches Netz zu dem Satz “ <i>Eine Orange wird manchmal auch Apfelsine genannt.</i> ”	170
5.33. Pseudocode, um inkonsistente Relationen in der Hypothesenwissensbasis <i>HKB</i> zu entdecken	176
5.34. Beweis des Theorems 5.5.1 durch Widerspruch	177
5.35. Beweis der Asymmetrie von in SUB eingebetteten Elementen durch Widerspruch	179
5.36. Gemäß Annahme 1 widersprüchliches Beispiel	180
5.37. Grafische Illustration des Beweises von Theorem 5.5.3	182
5.38. Eintrag <i>Festplatte</i> aus Wiktionary (Stand: August 2010)	185
5.39. Eintrag <i>Motor</i> aus Wiktionary (Stand: August 2010)	186
5.40. GUI des Annotationswerkzeuges SemChecker	191
5.41. Grafisches Benutzerinterface zur Anzeige des Resultates der logischen Validierung	191
6.1. Architektur der Entailmentextraktion von SemQuire	196
6.2. Auszug aus der Ankerliste zur Extraktion von Entailments	197
6.3. Datenmodell, um Entailments in der Datenbank abzulegen	199
7.1. Präzision der extrahierten Hyponymhypothesen für verschiedene Konfidenzwertintervalle (SQ=SemQuire, C=Cimiano)	214
D.1. Ankerliste zur Extraktion von Entailments	279
E.1. Lesbarkeitswertberechnung von DeLite	284
E.2. Screenshot der Benutzerschnittstelle von DeLite, bei dem eine Pronomenambiguität angezeigt wird für den Satz “ <i>Dr. Peters lädt Herrn Müller zum Essen ein, da er heute Geburtstag hat.</i> ”	290
E.3. Architektur der Relationsextraktion von SemDupl	291

Tabellenverzeichnis

3.1. Konfusionsmatrix	35
4.1. Ähnlichkeitswerte für Binärvektoren aus Manning/Schütze [MS99]	41
4.2. Ähnlichkeitsmaß für kontinuierliche Werte	41
4.3. Korrelationen der Ähnlichkeitsmaße mit einer menschlichen Beurteilung (M&C: Datensatz von Miller und Charles [MC91], R&G: Datensatz von Rubenstein und Goodenough [RG65])	44
4.4. Beispiel-Term-Dokument-Matrix für die formale Fuzzy-Begriffsanalyse	56
4.5. Von Girju, Badulescu und Moldovan [GBM06] vorgeschlagene Muster zur Extraktion von Meronymen	59
5.1. Eine Auswahl tiefer Regeln zur Extraktion von Hyponymen	86
5.2. Bestimmung der korrekten Subordinationsunterrelation	144
5.3. Ausgewählte MultiNet-Axiome	146
5.4. Auswahl von tiefen Meronym-Extraktionsregeln im MultiNet-Formalismus	147
5.5. Auswahl der korrekten Meronym-Unterrelation	162
5.6. Flache Extraktionsregeln zur Extraktion von Synonymen	168
5.7. Tiefe Extraktionsregeln zur Extraktion von Synonymiehypothesen	169
5.8. Logische Axiome, um einen Widerspruch herzuleiten	182
5.9. Unterschiede zwischen den ontologischen Sorten und semantischen Merkmalen der Institutionslesart von <i>bank.1.1</i> zu den verschiedenen Lesarten und Facetten von <i>Unternehmen</i>	189
5.10. Unterschiede zwischen den ontologischen Sorten und semantischen Merkmalen von <i>bank.2.1</i> zu den verschiedenen Lesarten und Facetten von <i>Unternehmen</i>	189
7.1. Eine Auswahl automatisch gelernter tiefer Regeln zur Extraktion von Hyponymen	209

7.2. Präzision der Hyponymhypothesen, die durch die Anwendung der gelernten Extraktionsregeln extrahiert wurden.	209
7.3. Korrelation einer Auswahl von Merkmalen zur Korrektheit der Relationshypothesen	214
7.4. Präzision der extrahierten Hyponymiehypothesen für unterschiedliche Konfidenzwertintervalle	214
7.5. Präzision der extrahierten Hyponymiehypothesen für verschiedene Konfidenzwertintervalle bei dem Verfahren von Cimiano et al. [CPSTS05]	214
7.6. Konfusionsmatrix für die Validierung von Hyponymiehypothesen	215
7.7. Akkuratheit, F-Wert, Präzision und Recall der Hyponymvalidierung	215
7.8. F-Wert, Präzision und Recall für die Bestimmung der korrekten Hyponymie-Unterrelation	215
7.9. Fehlerverteilung für extrahierte Subordinationshypothesen . . .	216
7.10. Gelernte Meronymextraktionsregeln	219
7.11. Präzisionswerte für gelernte Meronymextraktionsregeln	219
7.12. Konfusionsmatrix für GermaNet, Costello und SemQuire	220
7.13. Akkuratheit, F-Wert, Präzision und Recall für GermaNet, Costello und SemQuire	220
7.14. Konfusionsmatrix für die Validierung des semantikbasierten Meronymfilters	221
7.15. Akkuratheit, F-Wert, Präzision und Recall für die semantikbasierte Validierung von Meronymen	221
7.16. Korrelation einer Auswahl von Merkmalen mit der Hypothesenkorrektheit	225
7.17. F-Wert, Präzision und Recall für die automatische Bestimmung von Unterrelationen von MERO	225
7.18. Ausgewählte MultiNet-Axiome und die Anzahl der Beweise, bei denen diese angewendet werden.	226
7.19. Fehlerverteilung für extrahierte Meronymhypothesen	226
7.20. Konfusionsmatrix für die Validierung von Synonymen	229
7.21. Akkuratheit, F-Wert, Präzision und Recall für die Validierung von Synonymen	229
7.22. Konfusionsmatrix für den Semantik-basierten Validierungsfilter	229
7.23. Akkuratheit, F-Wert Präzision und Recall für den semantischen Validierungsfilter	230

7.24. Korrelation der Merkmale zur Hypothesenkorrektheit 230

7.25. Fehlerverteilung für extrahierte Synonymhypothesen 230

7.26. Eine Auswahl von angewandten Axiomen 235

7.27. Konfusionsmatrix für die extrahierten Entailmenthypothesen . . 237

7.28. Evaluationsmaße für die extrahierten Entailmenthypothesen . . 238

7.29. Auswahl von extrahierten Entailments 239

A.1. Tiefe Regeln zur Extraktion von Subordinationsbeziehungen . . 250

A.2. Flache Regeln zur Extraktion von Subordinationsbeziehungen . 252

A.3. Tiefe Regeln zur Extraktion von Meronymen 255

A.4. Flache Regeln zur Extraktion von Meronymen 257

B.1. Bei der Relationsextraktion eingesetzte Axiome 260

C.1. Hyponymhypothesen mit dem höchsten Konfidenzwert 262

C.2. Hyponymhypothesen sortiert nach ansteigendem Konfidenzwert 264

C.3. Meronym-Hypothesen mit dem höchstem Konfidenzwert 266

C.4. Meronymhypothesen sortiert nach ansteigendem Konfidenzwert 272

C.5. Synonym-Hypothesen mit dem höchstem Konfidenzwert 274

C.6. Synonymhypothesen sortiert nach aufsteigendem Konfidenzwert 276

1. Einleitung

1.1. Motivation

Eine große Wissensbasis ist eine Voraussetzung für eine Vielzahl von Anwendungen im Bereich der automatischen Sprachverarbeitung, wie Frage-Antwort- oder Information-Retrieval-Systeme. Bei einem Frage-Antwort-System geht es beispielsweise darum, eine Frage des Benutzers automatisch zu beantworten. Gewöhnlich verwendet das System dazu eine gegebene Textsammlung. In vielen Fällen ist die Frage anders formuliert als die Textstelle, die die richtige Antwort enthält. Um trotzdem die richtige Antwort zu finden, ist eine große Menge von Weltwissen erforderlich. Ein Mensch hat sich das dazu erforderliche Wissen im Laufe seines Lebens angeeignet. Einem Computer dagegen muss dieses Wissen explizit mitgeteilt werden.

Um eine Wissensbasis, die eine ausreichende Abdeckung besitzt, manuell zu erstellen, ist ein beträchtlicher Arbeitsaufwand nötig. Weiterhin sind Experten in Linguistik, Künstlicher Intelligenz und Wissensrepräsentation erforderlich, die unter Umständen nicht in dem erforderlichen Umfang verfügbar sind (Flaschenhals bei der Wissensakquisition) [Wat86, Seite 182]. Aus diesen Gründen sind automatische Verfahren hierbei von großer Bedeutung.

Die Sprachverarbeitungsverfahren zur Wissensextraktion (auch Sprachverarbeitungsverfahren im Allgemeinen) können bezüglich der verwendeten linguistischen Repräsentationsformen in verschiedene Ebenen unterteilt werden. Die flachste Ebene ist die Oberflächenrepräsentation. Verfahren, die auf dieser Ebene angesiedelt sind, verwenden keinerlei Parsing und werden als flache Verfahren bezeichnet. Die nächsttiefere Ebene ist die Chunkrepräsentation, wo einzelne Konstituenten ohne Berücksichtigung ihrer inneren Struktur erkannt werden. Ein Chunk bezeichnet dabei eine Phrase, eine inhaltlich deutbare Gruppe von Wörtern. Beispiele für semi-tiefe Repräsentationsstrukturen sind Abhängigkeitsbäume und Konstituentenbäume, die den Satz vollständig in einer syntaktischen Struktur abbilden, aber weder nach Lesarten unterschei-

den noch semantische Relationen enthalten. Die tiefste Struktur ist die tiefe semantische Repräsentation, die die Bedeutung eines Satzes auf einer logischen Ebene darstellt und auf Begriffen statt auf Wörtern basiert, d. h., es wird nach Lesarten unterschieden. Beispiele für eine solche Repräsentation sind Prädikatenlogik erster oder höherer Stufe, semantische Netze [Hel08] sowie Frames [Min74].

Die existierenden Verfahren zur Wissensakquisition sind überwiegend flach. Es existieren zwar einige semi-tiefe syntaktische Verfahren aber praktisch keine, die eine tiefe semantische Struktur verwenden. Zudem wird häufig auf Wörtern und nicht auf Begriffen basierendes Wissen extrahiert. Solches Wissen kann häufig zu fehlerhaften Inferenzen führen. Betrachte man beispielsweise die Fakten: “*Ein Schloss ist ein Teil der Tür.*” und “*Schloss Sanssouci ist ein Schloss.*” Ohne die Unterscheidung nach Lesarten von Wörtern könnte man ableiten, dass *Schloss Sanssouci* Teil einer Tür ist.

Praktisch allen vorhandenen Verfahren zur Wissensakquisition fehlt ein einheitliches Gebäude mit tiefem Parser, automatischem Beweiser und Anbindung an ein semantisches Lexikon, was für die Validierung und auch für die Extraktion des Wissens von großer Bedeutung ist. Auch von einer Wissensbasis und Axiomen machen diese Verfahren keinen Gebrauch. Durch baumbasierte syntaktische oder oberflächenbasierte Repräsentationen kann zudem die Bedeutung von Sätzen nicht formal adäquat beschrieben werden.

Im Folgenden wird ein Verfahren vorgestellt, das auf einer tiefen semantischen Repräsentation in Form semantischer Netze, die mithilfe eines linguistischen Parsers automatisch aus einem Textkorpus generiert werden, aufbaut. Es verwendet außerdem zur Validierung und Extraktion ein umfangreiches semantisches Lexikon und einen automatischen Beweiser mit Axiomen. Auf diese Weise können die Genauigkeit und der Umfang des extrahierten Wissens gegenüber flachen Ansätzen deutlich gesteigert werden.

1.2. Arten von Wissens

Eine Wissensbasis enthält typischerweise Repräsentationen folgender Arten von Wissen:

- lexikalisches Wissen
- syntaktisches Wissen
- semantisches Wissen

Im Folgenden werden einige Beispiele für diese Wissensarten untersucht. Das lexikalische Wissen enthält beispielsweise die Wortarten der im Satz vorkommenden Wörter. Sowohl dem lexikalischen als auch dem semantischen Wissen zugehörig sind Named-Entities. Der Subkategorisierungsrahmen, der die Argumente eines Verbs oder Nomens spezifiziert, gehört zum syntaktischen Wissen. Werden die Argumente um semantische Informationen angereichert, spricht man vom Valenzrahmen. Zu diesen semantischen Informationen gehören beispielsweise semantische Merkmale oder ontologische Sorten [Hel08]. Zudem sind der semantischen Ebene semantische Relationen wie Hyperonymie / Hyponymie (Ober- und Untertyp), Antonymie und Meronymie/Holonymie (Teil-Ganzes-Beziehungen) zugeordnet. Die Synonymie-Relation dagegen zählt zum lexikalischen Wissen. Neben semantischen und lexikalischen Relationen sind auch Entailments von großer Bedeutung für eine Wissensbasis. Ein *Entailment* ist in der Logik eine Folgerungsbeziehung zwischen zwei Formeln. In der natürlichen Sprache wird diese Relation meist nicht ganz so streng gehandhabt wie in der Logik. Man redet von einem Textual-Entailment zwischen zwei Textpassagen T_1 und T_2 , wenn, falls die Aussage aus Text T_1 erfüllt ist, auch die Aussage aus Text T_2 fast sicher erfüllt ist. Textual-Entailments sind auf der Oberfläche definiert, während semantische Entailments auf der Bedeutungsrepräsentation von Sätzen aufbauen.

1.3. Vorteile der Nutzung von Wissen bei der natürlichsprachlichen Verarbeitung

Dieser Abschnitt soll einen kurzen Überblick geben, inwieweit eine große Wissensbasis bei der automatischen Sprachverarbeitung von Nutzen sein kann.

Das Wissen über den Valenzrahmen wird u. a. von linguistischen Parsern verwendet, um Mehrdeutigkeiten in der syntaktischen Struktur oder mehrdeutige Wörter (Polysemie) zu disambiguieren (Word-Sense-Disambiguation) [Har03].

Semantische und lexikalische Relationen sind für alle Bereiche der automatischen Sprachverarbeitung von großer Bedeutung, die Inferenzen beinhalten, wie z. B. für die Erkennung von Textual-Entailments [RDH⁺83, BM06]. Auf Anwendungsebene können sie für Frage-Antwort- oder Information-Retrieval-Systeme eingesetzt werden. Für Frage-Antwort-Systeme soll dies im Folgenden genauer ausgeführt werden.

Ein Frage-Antwort-System ist ein System, das in natürlicher Sprache gestellte Fragen eines Benutzers beantwortet. Normalerweise wird dazu ein gegebener Textkorpus (ggf. unterstützt durch weiteres Hintergrundwissen) verwendet.

Als Beispiel betrachte man folgende Frage: “*Wer ist der Chef der Daimler AG?*” Angenommen, der Textkorpus enthalte die Information, dass “*Dieter Zetsche der Vorstandsvorsitzende der Daimler AG*” ist. Wenn nun zusätzlich bekannt ist, dass *Vorstandsvorsitzender* ein Hyponym von *Chef* ist, kann die Frage korrekt mit *Dieter Zetsche* beantwortet werden.

Ein Beispiel, bei dem Meronymiebeziehungen hilfreich sein können, ist die folgende Frage: “*Welche aktuellen Fußballnationalspieler sind in Bayern geboren?*” Man nehme an, in der Datenbank des Frage-Antwort-Systems befinden sich folgende Sätze: “*Philipp Lahm wurde in München geboren. Philipp Lahm ist ein aktueller deutscher Fußballnationalspieler.*” Wenn das System weiterhin die Information besitzt, dass München ein Meronym von Bayern ist, kann die Frage korrekt mit *Philipp Lahm* (und eventuell den Namen weiterer Spieler, die in Bayern geboren sind) beantwortet werden.

Auch Synonyme sind von großer Bedeutung für Frage-Antwort-Systeme. Betrachte man die Frage: *Wer ist der Präsident der USA?* Ein Text in der Datenbank des Frage-Antwort-Systems enthalte den Satz: *Barack Obama ist der Präsident der Vereinigten Staaten von Amerika.* Wenn nun bekannt ist, dass *USA* ein Synonym von *Vereinigte Staaten von Amerika* ist, dann kann die Frage mithilfe dieses Textes beantwortet werden.

Entailments sind ebenfalls für Frage-Antwort-Systeme von Bedeutung. Angenommen, der Benutzer stellt die Frage: “*Wann wurde Angela Merkel geboren?*” Ein Text der Korpus enthalte den Satz: “*Der Geburtstag von Angela Merkel ist der 17. Juli 1954.*” Man nehme an, in der Wissensbasis ist das bidirektionale Entailment enthalten: *Der Geburtstag von X ist Y* \Leftrightarrow *X ist am Y geboren.*, wobei das Entailment in einer Oberflächendarstellung wie angegeben oder in einer logischen Repräsentation, wie beispielsweise Prädikatenlogik, repräsentiert sein kann. In diesem Fall kann die Frage mithilfe des Entailments mit dem in der Wissensbasis enthaltenen Text in Bezug gebracht und somit die richtige Antwort *17. Juli 1954* ermittelt werden.

Eine Beispielfrage, bei der sich ein gerichtetes Entailment nützlich erweisen kann, ist “*Arbeitet Dieter Zetsche noch bei der Daimler AG?*”. Man nehme an, der Text enthalte die Information: “*Dieter Zetsche ist Vorstandsvorsitzender der Daimler AG*”. Wenn nun ein Entailment in der Wissensbasis existiert, dass jemand, der Vorstandsvorsitzender eines Unternehmens ist, auch dort arbeitet,

dann kann diese Frage mit *ja* beantwortet werden.

Darüber hinaus ist eine gut ausgebaute Wissensbasis aber auch in vielen anderen Bereichen der automatischen Sprachverarbeitung von Vorteil. Beispielsweise kann ein Textgenerierungssystem, um Wortwiederholungen zu vermeiden, einen Begriff durch ein Synonym oder Hyperonym ersetzen [KD96]. Ein weiterer Anwendungsbereich ist das Erkennen von Duplikaten oder Beinahe-Duplikaten. So könnte ein Plagiator, um das Plagiat zu verschleiern, Wörter durch Synonyme, Hyperonyme, Hyponyme, Meronyme oder Holonyme austauschen. Zudem wären auch Umformulierungen ganzer Sätze durch Paraphrasen oder Textual-Entailments denkbar.

Semantische Relationen werden auch verwendet, um Bridging-Referenzen aufzulösen [GHH06a], was notwendig für die Assimilation (siehe Glossar in Abschnitt F) semantischer Netzwerke ist. Auch bei der Anaphernresolution ist eine Taxonomie hilfreich, um z. B. zu erkennen, auf welchen Antezedenten in einem Satz sich ein Ausdruck beziehen kann. Betrachtet man das Beispiel: *“Das Haus wurde schließlich abgerissen. Laut dem Baudezernenten war dieses Gebäude schon lange baufällig.”* Durch Verwendung einer Taxonomie kann man erkennen, dass sich der Ausdruck *dieses Gebäude* auf das *Haus* im Satz davor beziehen muss.

Die Verwendung von Wissen kann allerdings auch zu Problemen führen. So können Inkonsistenzen in der Wissensbasis dazu führen, dass sich jede Aussage ableiten lässt. Zudem ist es möglich, dass durch einige sehr zentrale fehlerhafte Fakten, riesige Mengen von inkorrekten Aussagen folgerbar ist. Im Prinzip können sich dadurch die Resultate des Sprachverarbeitungssystems auch verschlechtern. Wichtig bei der Verwendung von Wissen ist daher eine gute Validierung dieses Wissens. Ein weiteres Problem besteht in dem Aufwand, der nötig ist, um eine große Menge von Wissen zu erzeugen. Dabei hilfreich kann eine automatische Wissensakquisition sein, wie im nächsten Abschnitt beschrieben.

1.4. Vorteile einer automatischen Wissensakquisition

Die manuelle Erzeugung von großen Wissensbasen ist mit einem immensen Arbeitsaufwand verbunden. Zudem sind zur manuellen Erstellung von Wissensbasen meist Domänenexperten (beispielsweise Mediziner, Chemiker, Physiker,

etc.) erforderlich, die evtl. nicht im gewünschten Maße verfügbar sind oder die nicht mit Wissensrepräsentationsformalismen sowie semantischen und lexikalischen Relationen vertraut sind. Daher wurde eine Vielzahl von Methoden entwickelt, das Wissen automatisch aus großen Textbeständen zu extrahieren. Große Textbestände zu erhalten, ist durch die zunehmende elektronische Speicherung von Texten sowie durch die Entwicklung des Internets sehr einfach geworden. Automatische Verfahren sind zudem meist leicht auf andere Domänen übertragbar. So muss man in vielen Fällen lediglich einen domänenspezifischen Korpus als Eingabe für ein automatisches Verfahren angeben und dieses Verfahren darauf neu anwenden, um eine neue Ontologie aufzubauen. Würde man dagegen die Wissensbasis manuell aufbauen, kann bei der Betrachtung neuer Domänen meist nur in sehr geringem Umfang Wissen von anderen Domänen wiederverwendet werden.

Allerdings erzeugen automatische Verfahren neben korrekten Wissenshypothesen normalerweise auch eine große Anzahl fehlerhafter Hypothesen. Das bedeutet, dass Validierungsverfahren entwickelt werden müssen, um solche Hypothesen herauszufiltern oder um einen Konfidenzwert zu berechnen, wobei idealerweise den korrekten Hypothesen ein sehr hoher und den fehlerhaften Hypothesen ein sehr geringer Wert zugewiesen werden sollte.

1.5. Einordnung der Arbeit

1.5.1. Thematische Einordnung der Arbeit

Thematisch ist diese Arbeit in die Bereiche der automatischen Sprachverarbeitung, des maschinellen Lernens und der Logik eingeordnet. Im Bereich der Sprachwissenschaft hat diese Arbeit Berührungspunkte mit Semantik (z. B. Paraphrasierungen), Computerlexikografie und Korpuslinguistik. Die hier eingesetzten maschinellen Lernverfahren konzentrieren sich in erster Linie auf graphbasierte Methoden, wie den Einsatz eines Graphkernels und das Lernen von Graphstrukturen. Logische Methoden werden verwendet um die Ontologie zu validieren und um Wissen zu extrahieren. Zudem wird das extrahierte Wissen in einem logischen Wissensrepräsentationsformalismus abgelegt.

1.5.2. Organisatorische Einordnung der Arbeit

Organisatorisch wurde diese Arbeit im Rahmen des DFG-Projektes *SemDupl: Semantische Duplikaterkennung mithilfe von Textual-Entailment* (HE 2847/11-1) durchgeführt (siehe Abschnitt E.2), in dem es das Ziel war, Duplikate mithilfe einer tiefen semantischen Analyse unter Verwendung einer umfangreichen Wissensbasis zu identifizieren.

Das dieser Arbeit zugrunde liegende System verwendet viele der am Lehrgebiet *Intelligente Informations- und Kommunikationssysteme* entwickelten Softwarekomponenten, insbesondere

- Automatische Beweiser: Der automatische Beweiser *MultiNet-Beweiser* [Glö08] wird verwendet, um Wissen zu extrahieren. Der Beweiser *EKR-Hyper* [BFP07] kam für die Validierung zum Einsatz. Diese zwei Schritte sind weitgehend unabhängig voneinander realisiert.
- Tiefer Parser WOCADI¹ [Har03]: Die hier verwendeten Wissensakquisitionsverfahren setzen nicht direkt auf einer Oberflächenrepräsentation, sondern auf tiefen semantischen Strukturen in Form semantischer Netzwerke, auf. Um diese automatisch aus Text zu erzeugen, kommt der semantisch-syntaktische Parser WOCADI zum Einsatz.

Neben Anwendungssystemen werden auch verschiedene an oben erwähntem Lehrgebiet entwickelte Ressourcen verwendet. Diese sind:

- HaGenLex² [HHO03]: Das semantische Lexikon HaGenLex wird zur Wissensvalidierung herangezogen und zur Bestimmung der korrekten Unterrelation.
- Semantische Netzwerke: Aus verschiedene Textkorpora erzeugte semantische Netzwerke bilden die Grundlage für die Wissensakquisition.
- Axiome: Axiome werden zur Wissensakquisition und zur Wissensvalidierung verwendet. Ein Teil der Axiome stammt aus dem Buch von Helbig [Hel08], ein Teil wurde von Ingo Glöckner entwickelt und ein weiterer Teil wurde speziell für das hier beschriebene System SemQuire entworfen.

Eine genauere Beschreibung der eingesetzten Ressourcen erfolgt in Abschnitt 3.2.

¹WOCADI ist die Abkürzung für “**W**ord-**c**lass based **d**isambiguating parser”

²HaGenLex ist die Abkürzung für **H**agen **G**erman **L**exicon

1.6. Zusammenfassung und Thesen

In dieser Arbeit wird ein Verfahren (SemQuire) beschrieben, das automatisch Wissen (semantische und lexikalische Relationen sowie Entailments) aus Texten extrahiert, wofür neben der Verwendung von linguistischem Wissen, Statistik (insbesondere maschineller Lernverfahren) logische Inferenzen eingesetzt werden. Die Texte werden dazu mithilfe eines tiefen linguistischen Parsers namens WOCADI in semantische Netzwerke gemäß dem MultiNet-Paradigma³[Hel08] überführt. Aus diesen semantischen Netzen wird anschließend das Wissen extrahiert. Das extrahierte Wissen ist dabei begriffs- und nicht wortbasiert. Bezüglich des beschriebenen Verfahrens SemQuire werden in dieser Arbeit dabei die folgenden Thesen vertreten:

A Gesamterfolg

1. Flache Verfahren sind ungeeignet für die zuverlässige Wissensextraktion. Durch die Basierung des Wissens auf Wörtern anstelle von Begriffen entstehen häufig Fehler. Zudem ist eine Validierung von solchem Wissen mithilfe logischer Verfahren nur sehr eingeschränkt möglich.
2. Durch die Verwendung einer tiefen semantischen Repräsentation können die Qualität des Wissens und die Quantität von qualitativ hochwertigem Wissen deutlich gesteigert werden im Vergleich mit einem rein flachen Ansatz.

B Anwendbarkeit

1. Der hier vorgestellte Ansatz ist auf Texte aus beliebigen Domänen anwendbar.
2. Der hier vorgestellte Ansatz wird zwar auf ein Lexikon (Wikipedia) angewendet, ist aber unabhängig von der Textart.

C Beitrag der Komponenten

1. Durch die Verwendung eines einheitlichen Gebäudes auf konzeptueller Ebene und in Bezug auf die eingesetzten Systeme kann der Wissensextraktionsprozess sehr leicht auf andere Sprachen portiert werden.
2. Nur dadurch, dass der ganze Wissensextraktionsprozess in ein konzeptuelles Gebäude eingebettet ist (hier: eine MultiNet-Wissensbasis, eine Liste von Axiomen sowie ein semantisches Lexikon), ist eine zuverlässige und umfangreiche Wissensextraktion möglich.

³MultiNet ist die Abkürzung für “**M**ultilayer **E**xtended **S**emantic **N**etworks”

3. Nur durch die Verwendung eines einheitlichen Gebäudes von Anwendungssystemen, das den automatischen Beweiser sowie den tiefen Parser WOCADI einschließt, ist eine zuverlässige und umfangreiche Wissensextraktion möglich.
4. Durch Verwendung eines automatischen Theorembeweislers sowohl zur Extraktion von Relationen als auch zur logischen Validierung kann sowohl die Qualität von Hypothesen als auch die Quantität von Hypothesen mit guter Qualität gesteigert werden.
5. Der Einsatz zusätzlicher lexikalischer Ressourcen, beispielsweise neue oder modifizierte Lexikoneinträge, können die Validierung in vielen Fällen automatisch verbessern.

D Hybrid-Anteil

1. Auch ein eher flaches, auf Tokenlisten (zur Erläuterung von Tokens siehe das Glossar in Abschnitt F) basierendes Verfahren kann durch Verwendung des MultiNet-Parsers WOCADI semantische Informationen ausnutzen und dadurch ebenfalls auf Begriffen basierendes Wissen extrahieren. Die Verwendung von flachen Verfahren ist notwendig, um auch aus Sätzen, die nicht parsbar sind, Wissen zu extrahieren.
2. Die hier vorgestellten Validierungskomponenten können auch für Wissen verwendet werden, das durch ein eher flaches, auf Tokenlisten basierendes Verfahren extrahiert wurde.

E Bootstrapping-Anteil

1. Die verschiedenen extrahierten Wissensarten können wechselseitig für die Validierung verwendet werden. Beispielsweise lässt sich durch zusätzliche Hyponyme automatisch die Qualität der Meronymhypothesen verbessern.

1.7. Aufbau der Arbeit

Bevor weiter ins Detail gegangen wird, soll die Kapiteleinteilung dieser Arbeit angegeben werden. Kapitel 2 beschreibt die Arten von Wissen, die in einer Wissensbasis vorhanden sein sollten. Kapitel 3 enthält einen Überblick über den MultiNet-Formalismus zur Wissensrepräsentation, auf dem in dieser Arbeit aufgebaut werden soll. Kapitel 4 erläutert verschiedene Methoden der automatischen Wissensextraktion, wie sie dem bisherigen Stand der Forschung

entsprechen. In Kapitel 5 erfolgt eine Beschreibung der im Rahmen dieser Arbeit entwickelten Verfahren zum Wissensaufbau von semantischen und lexikalischen Relationen. In Kapitel 6 wird die Entailmentextraktion erläutert. Kapitel 7 enthält die Evaluation der in Kapitel 5 beschriebenen Verfahren. Kapitel 8 gibt einerseits eine Zusammenfassung der in dieser Arbeit erbrachten Leistungen, beschreibt aber auch mögliche Weiterentwicklungen. Der Anhang enthält schließlich die zur Wissensextraktion verwendeten Extraktionsregeln (siehe Kapitel 8.3) und Axiome (siehe Kapitel B), Listen extrahierter Hypothesen (siehe Kapitel C), listet die zur Entailmentextraktion verwendete Ankerliste auf (siehe Kapitel D) und gibt einen Überblick über weitere auf MultiNet basierende Anwendungssysteme (siehe Kapitel E), die Gebrauch von einer Wissensbasis machen wie das Lesbarkeitsbeurteilungssystem DeLite, das Duplikatserkennungssystem SemDupl und das Frage-Antwort-System LogAnswer. Schließlich enthält der Anhang ein Glossar mit in dieser Arbeit verwendeten Definitionen (siehe Kapitel F).

1.8. Typografische Konventionen

In dieser Arbeit werden unterschiedliche Arten von Textinhalten auf verschiedene Weise dargestellt, was die Lesbarkeit verbessern soll.

Natürlichsprachliche Ausdrücke, die in Beispielen auftauchen, werden in Kursivschrift dargestellt. Mathematische Variablen, Konstanten und Prädikate werden ebenfalls in kursiver Schrift dargestellt, wobei Mengen und Prädikate großgeschrieben, Variablen (ausgenommen Zufallsvariablen) und Konstanten dagegen kleingeschrieben werden. MultiNet-Relationen und -Funktionen werden mit Kapitälchen gesetzt (z. B. SUB). Begriffe werden entweder durch einen Zahlensuffix (z. B. 1.1) oder durch eckige Klammern ($\langle \rangle$) gekennzeichnet.

2. Typische Arten von Wissen

2.1. Synonymie

Als Synonymie bezeichnet man die Tatsache, dass unterschiedliche Ausdrücke die gleiche Bedeutung besitzen. Diese Ausdrücke sind in diesem Fall synonym zueinander [Lyo95]. Häufig werden damit Wörter in Beziehung gesetzt. Dies kann aber zu Problemen führen, wenn Wörter mehrere Lesarten besitzen. Betrachte man beispielsweise die Synonyme *Pferd* und *Gaul*. Es existiert eine Lesart von *Pferd*, die ein Sportgerät ist. Ohne Berücksichtigung von Lesarten lässt sich beispielsweise folgern, dass *Gaul* ein Sportgerät ist. Im Folgenden wird die Bezeichnung Synonymie nur für die Synonymie zwischen Begriffen verwendet. $\text{SYNO}(\text{bündnis.1.1}, \text{allianz.1.1})$ bedeutet beispielsweise, dass die Lesart 1.1 des Wortes *Bündnis* die gleiche Bedeutung hat wie Lesart 1.1 des Wortes *Allianz*. Grundsätzlich gilt: $\neg \text{SYNO}(\text{wort.} \langle \text{reading}_1 \rangle, \text{wort.} \langle \text{reading}_2 \rangle)$ mit $\text{reading}_1 \neq \text{reading}_2$, d. h., zwei verschiedene Lesarten desselben Wortes können nicht synonym zueinander sein. Eine genauere Beschreibung der MultiNet-Lesarten findet man in Abschnitt 3.1.

Die Synonymie zwischen Wörtern wird dagegen als Wortsynonymie bezeichnet und folgendermaßen definiert:

Definition 2.1.1 *Es besteht eine Wortsynonymie zwischen zwei Wörtern w_1 und w_2 , wenn mindestens eine Lesart von w_1 existiert, die synonym zu einer Lesart von w_2 ist.*

2.2. Subordination

Helbig fasst Hyponymie- und Instanzrelationen zur Subordination zusammen [Hel08]. Hyponymie ist gemäß Bußmann [Buß08] “der *Terminus für die semantische Relation der Unterordnung im Sinne einer inhaltsgemäßen Spezifizierung [...] Bei Ausdrücken, die eine Extension haben, ergibt sich die Hyponymie-Relation als Teilmengenbeziehung. L_1 ist ein Hyponym von L_2 genau dann,*

wenn die Extension von L_1 enthalten ist in der Extension von L_2 “ (MultiNet-Relation für Subordination: SUB0).

Beispielsweise ist *Hund* ein Hyponym von *Tier* und *Tier* ein Hyperonym von *Hund*. Es wird hier nicht der Definition von Lyons gefolgt [Lyo95], bei der die Synonymrelation bei der Hyponymrelation mit eingeschlossen ist, d. h., die Hyponymrelation ist asymmetrisch, sodass folgt: $\text{SUB0}(a, b) \rightarrow \neg\text{SUB0}(b, a)$.

Mit einer Instanzrelation wird angegeben, welchem Begriff eine individuelle Entität untergeordnet ist, d. h., eine Instanz ist Element der Extension des übergeordneten Begriffes. Beispielsweise ist *Deutschland* ein *Land*, d. h., die Instanz *Deutschland* ist in der Extension von *Land* enthalten.

Die Subordinationsrelation wird von Helbig [Hel08, Seite 529] in drei Unterrelationen unterteilt:

- Instanzrelation oder Hyponymie zwischen Begriffen, die keine Situationen oder Relationen repräsentieren. Diese Subordinationsunterrelation kommt in der Praxis am häufigsten vor.
Beispiel: *Hund* ist ein Hyponym von *Tier*. MultiNet-Relation: SUB
- Instanzrelation, Hyponymie oder Troponymie zwischen Situationen. Beispiel: *Hochzeitsfeier* ist ein Hyponym von *Feier*, wobei die Ausdrücke *Hochzeitsfeier* und *Feier* jeweils Situationen kennzeichnen. MultiNet-Relation: SUBS
- Instanzrelation oder Hyponymie zwischen Relationen.
Beispiel: *Geschwindigkeitsdifferenz* ist ein Hyponym von *Differenz*, wobei *Geschwindigkeitsdifferenz* und *Differenz* beides Relationen zwischen zwei Zahlen sind. MultiNet-Relation: SUBR

Da bei der automatischen Verarbeitung eine konkrete Subordinationsrelation aufgrund fehlenden Weltwissens nicht immer eindeutig einer der drei Unterrelationen zugeordnet werden kann, wurde von Helbig die MultiNet-Relation namens SUB0 eingeführt, die alle drei Unterrelationen umfasst.

Analog wie bei Synonymie wird die Bezeichnung Worthyponymie verwendet, die die Hyponymrelation auf Wörter überträgt. Dabei sollten mit Worthyponymen allerdings keine Inferenzen durchgeführt werden, da dadurch in vielen Fällen unsinnige Fakten abgeleitet werden können. Analog werden die Relationen *Wortmeronymie* und *Wortantonymie* verwendet.

2.3. Meronymie

Meronymie bezeichnet eine Teil-Ganzes- oder Element-Menge-Relation. Der Teil wird als Meronym des dieses enthaltenden Objektes, des sogenannten Holonyms, bezeichnet. Gemäß Winston kann die Meronymie-Relation in folgende Unterrelationen unterteilt werden [WCH87]:

- **Komponente-Gesamtheit:** eine Relation zwischen einem Objekt und einer seiner Komponenten. Laut Winston ist es dabei von Bedeutung, dass das Objekt und seine Komponente separat voneinander wahrgenommen werden können. Beispielsweise ist es möglich, ein Rad von dem dazugehörigen Auto zu unterscheiden. Diese ist auch in der Praxis die am häufigsten vorkommende Relation. Verwendete MultiNet-Relation: PARS
- **Element-Menge:** Diese Unterrelation repräsentiert die Mitgliedschaft in einer Menge. Beispiel: “*Ein **Fußballspieler** ist Teil einer **Fußballmannschaft**.*” MultiNet-Relation: ELMT
- **Maßeinheiten-Portionen:** Relationen, die Maßeinheiten-, Untereinheiten sowie Portionen betreffen. Beispiel: “*Ein **Stück Kuchen** ist Teil eines **Kuchens**.*” / “*Ein **Meter** ist Teil eines **Kilometers**.*” MultiNet-Relation: PARS oder aber TEMP für Zeiteinheiten.
- **Material-Objekt:** Diese Unterrelation repräsentiert die chemische Zusammensetzung eines Objektes. Beispiel: “***Alkohol** ist Bestandteil des **Weins**.*” / “***Sauerstoff** ist in **Luft** enthalten.*” / “*Der **Tisch** besteht aus **Holz**.*” MultiNet-Relation: PARS oder aber ORIGM^{-1} , falls es sich bei dem Holonym um ein diskretes Objekt und nicht um eine Substanz handelt. ORIGM^{-1} bezeichnet dabei die inverse Relation von ORIGM , d. h., $\forall x, y : \text{ORIGM}(x, y) \Leftrightarrow \text{ORIGM}^{-1}(y, x)$
- **Merkmal-Aktivität:** Aktivitäten können normalerweise in verschiedene Teilaktivitäten zerlegt werden. Beispiel: Die folgenden Teilaktivitäten gehören zu der Aktivität *Abendessen im Restaurant*: *Ein Restaurant besuchen, Bestellen, Essen* und *Bezahlen*. MultiNet-Relation: HSIT

Zusätzlich zu diesen Unterrelationen betrachtet Helbig [Hel08, Seite 530] die Teilmengenrelation als weitere Meronymie-Unterrelation (MultiNet-Relation: SUBM). Beispielsweise ist ein *Bataillon* eine Teilmenge und kein Element einer

Brigade. Tatsächlich sind sowohl die Elemente vom *Bataillon* als auch von *Brigade Soldaten*.

Da man bei der automatischen Verarbeitung eine konkrete Meronymierelation aufgrund fehlenden Weltwissens nicht immer eindeutig einer der obigen Unterrelationen zuordnen kann, hat Helbig zusätzlich als Hilfskonstrukt die MultiNet-Relation MERO eingeführt, die alle diese Unterrelationen zusammenfasst.

2.4. Entailments und Paraphrasen

Ein Textual-Entailment ist eine Relation zwischen zwei Textabschnitten von der Art, dass aus der Gültigkeit des einen Abschnittes die Gültigkeit des anderen fast sicher folgt. Eine Sonderform des Textual-Entailments ist die Paraphrase. Gemäß Fronkin, Rodman und Hymas sind zwei Texte Paraphrasen voneinander, wenn diese dieselbe Bedeutung haben [FRH02].

Beispiele für Entailments sind:

1. *Björn Borg besiegte im Wimbledonfinale 1980 John McEnroe.* \Leftrightarrow *John McEnroe verlor im Wimbledonfinale 1980 gegen Björn Borg.*
2. *Der Vater schenkte seinem Sohn eine Eisenbahn.* \Rightarrow *Der Sohn erhielt von seinem Vater eine Eisenbahn.*
3. *Herr Franck gefiel sein Urlaub in den Alpen sehr gut.* \Rightarrow *Herr Franck reiste in seinem Urlaub in die Alpen.*
4. *Herr Franck reiste in seinem Urlaub in die Alpen.* \Rightarrow *Herr Franck wohnt nicht in den Alpen.*
5. *Boris Becker siegte gegen Stephan Edberg im Finale von Wimbledon.* \Leftrightarrow *Beckers Sieg gegen Stephan Edberg in Wimbledon*

Im ersten Beispiel gilt das Entailment in beide Richtungen, d. h., man kann aus dem ersten Satz den zweiten Satz folgern und umgekehrt. In diesem Fall spricht man bei Verwendung von Oberflächenrepräsentationen auch von Paraphrasen.

Im zweiten Beispiel dagegen gilt das Entailment nur in eine Richtung. Aus der Tatsache, dass der Vater seinem Sohn eine Eisenbahn geschenkt hat, kann man folgern, dass der Sohn diese auch erhalten hat, nicht aber unbedingt

umgekehrt. Wenn nur der zweite Satz des Beispiels bekannt ist, könnte es unter Umständen auch sein, dass die Eisenbahn dem Sohn nur geliehen wurde.

Das dritte Beispiel ist eine Präsupposition. So macht die Aussage in dem ersten Satz des Beispiels nur Sinn, wenn Herr Franck tatsächlich in den Alpen war.

Das vierte Beispiel ist eine Implikatur. Aus der Aussage, dass Herr Franck für seinen Urlaub in die Alpen fährt, kann man folgern, dass er dort nicht wohnt.

Das fünfte Beispiel enthält eine Nominalisierung. Anhand der Tatsache, dass Becker gegen Edberg in Wimbledon gewann, lässt sich folgern, dass es einen Sieg Beckers gegen Edberg in Wimbledon gab und umgekehrt.

Alle die vier Arten sind für diese Arbeit relevant und sollen aus Texten extrahiert werden können.

Im Unterschied zu Textual-Entailments, die von praktisch allen aktuellen Entailmentlernverfahren erzeugt werden, wird in dieser Arbeit das Lernen von *semantischen Entailments* beschrieben. Diese setzen nicht auf der Oberflächenrepräsentation auf, sondern auf den Bedeutungsrepräsentationen der zu vergleichenden Sätze (hier in Form semantischer Netze).