Karlsruher Institut für Technologie

Masterarbeit

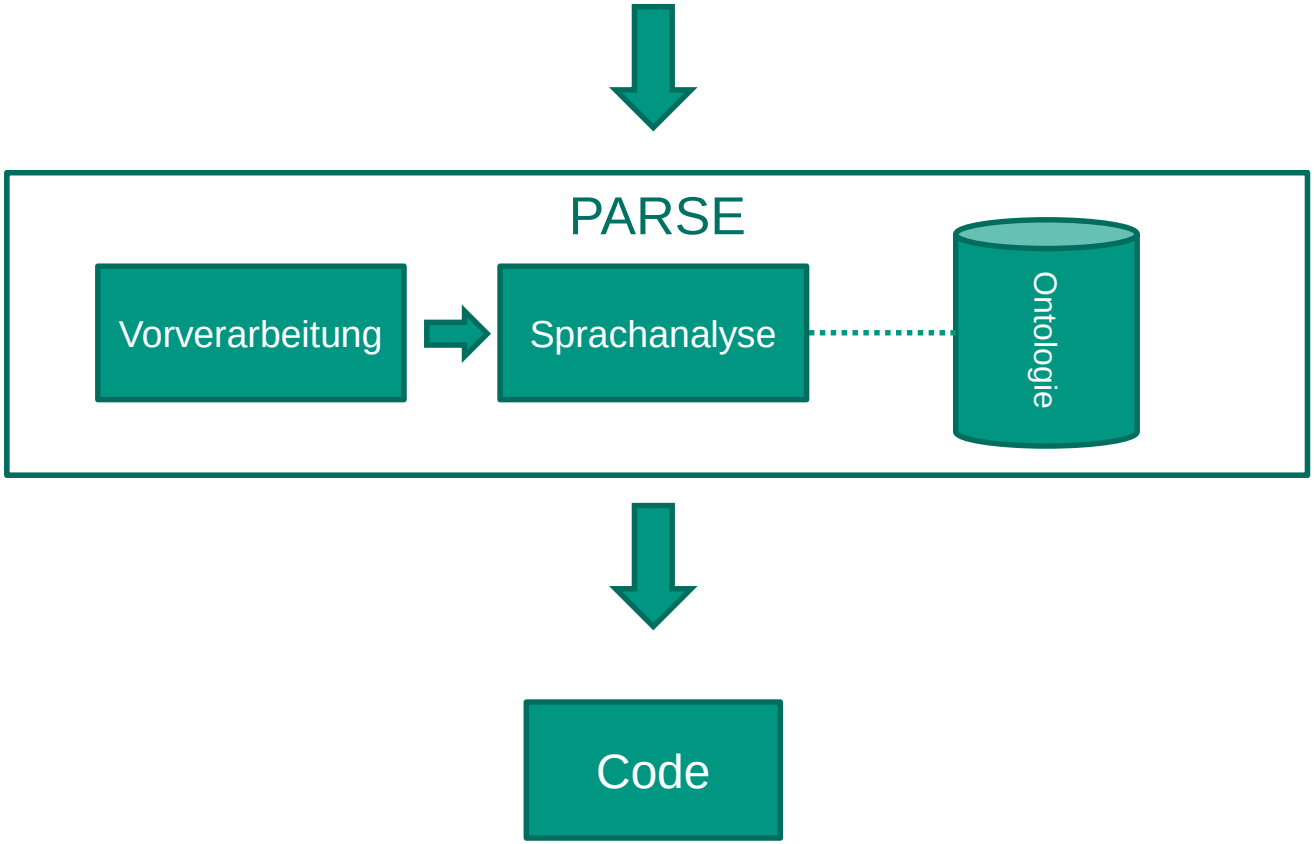
Themenextraktion zur Domänenauswahl für Programmierung in natürlicher Sprache

Jan Keim
Betreut von Sebastian Weigelt und Tobias Hey

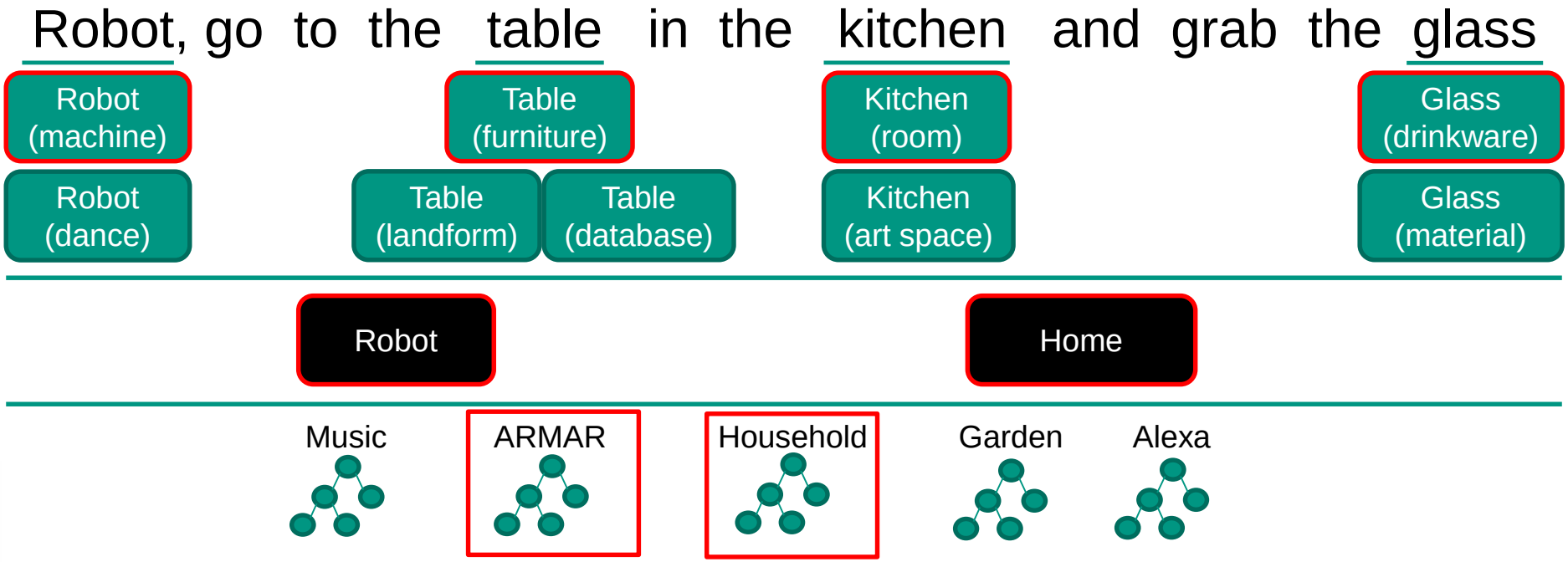
IPD Tichy, Fakultät für Informatik

Motivation

Robot, go to the table in the kitchen and grab the glass



Motivation



- Ziel: Themenextraktion
- Ansatz: Graphbasiert
  1. Auflösung der Mehrdeutigkeit von Nomen
  2. Graph erstellen und Themen bestimmen
- Anwendung: Auswahl von themenspezifischen Ontologien

Verwandte Arbeiten

- Verschiedene Arbeiten im Bereich der Themenextraktion

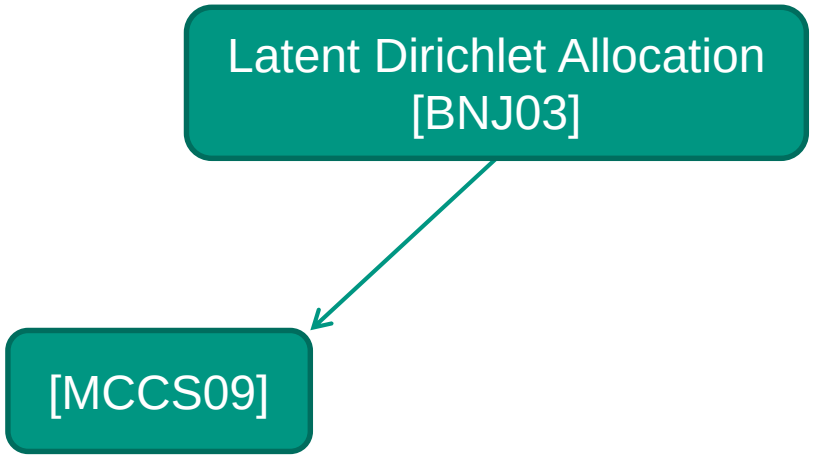
Latent Dirichlet Allocation [BNJ03]

[BNJ03] Blei et al.: Latent dirichlet allocation

- Statistisches Modell
- Zuordnung von Dokumenten zu k Themen

Verwandte Arbeiten

- Verschiedene Arbeiten im Bereich der Themenextraktion

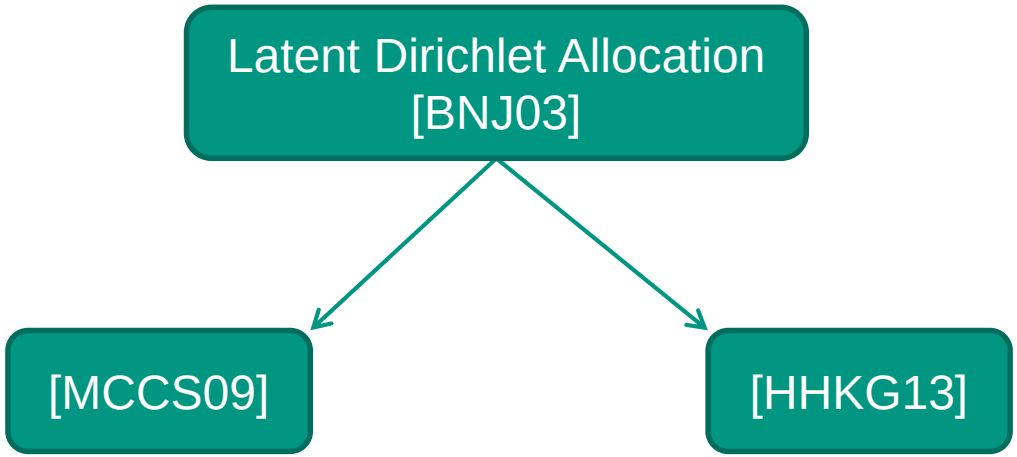


[MCCS09] Magatti et al.: Automatic labeling of topics

- Wichtige Begriffe aus LDA
- Konsens in Baumhierarchie

Verwandte Arbeiten

- Verschiedene Arbeiten im Bereich der Themenextraktion

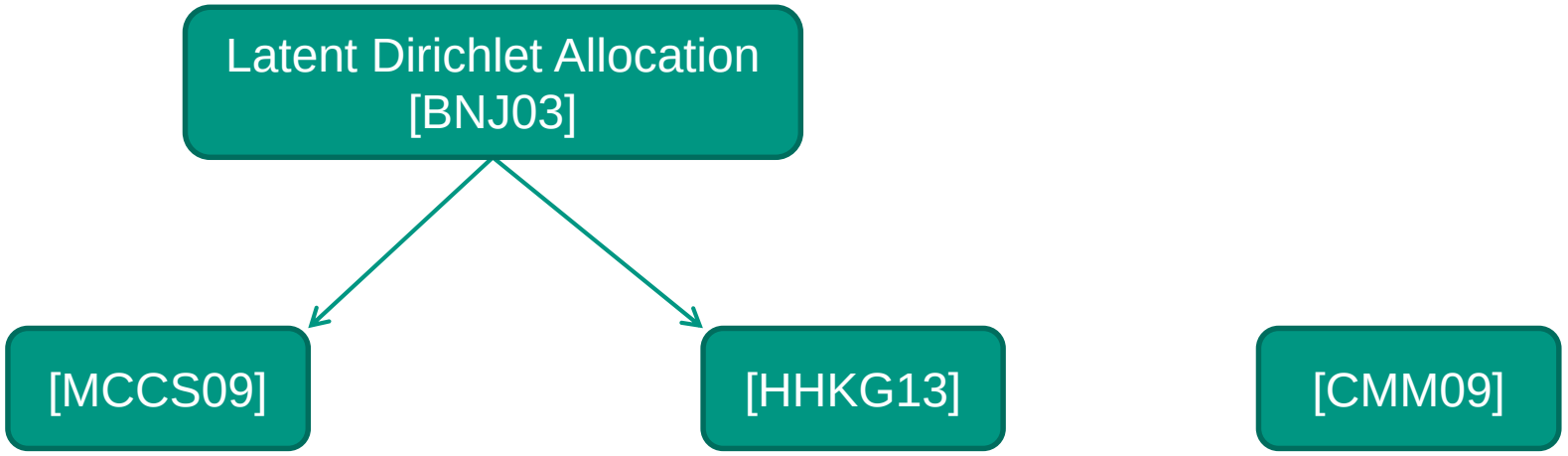


[HHKG13] Hulpus et al.: Unsupervised graph-based topic labelling using DBpedia

- Aus wichtigen Begriffen Themengraphen erstellen
- Zentrale Knoten im Themengraph sind Themen

Verwandte Arbeiten

- Verschiedene Arbeiten im Bereich der Themenextraktion

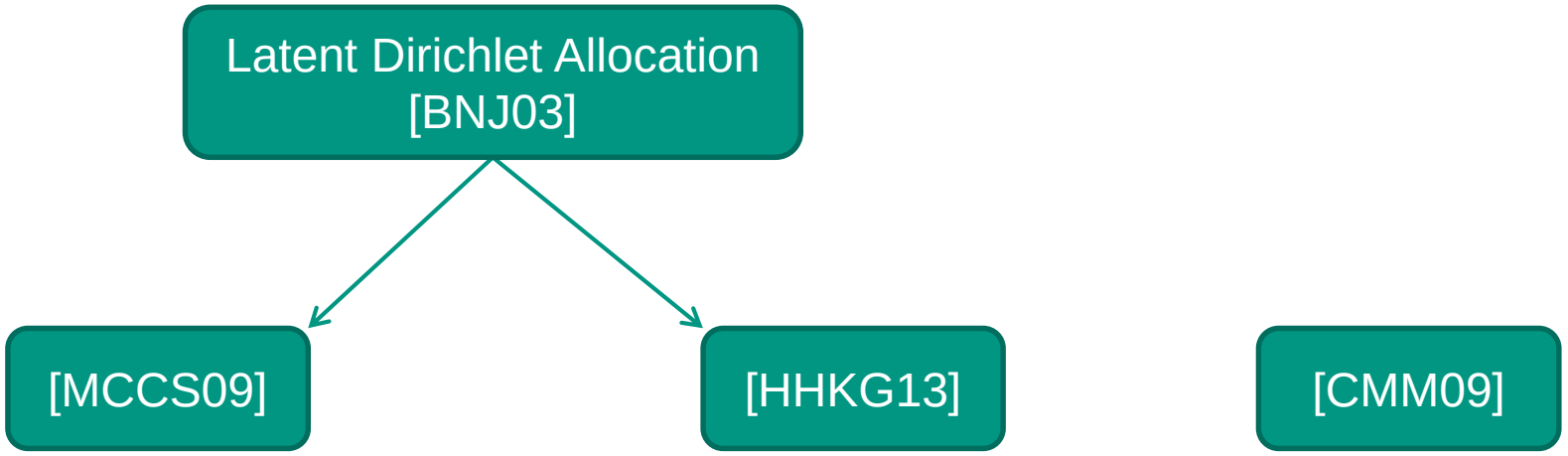


[CMM09] Coursey et al.: Using encyclopedic knowledge for automatic topic identification

- Graph aus Wikipedia
- Zuordnung wichtiger Begriffe zu Knoten im Graph
- Zentrale Knoten sind Themen

Verwandte Arbeiten

- Verschiedene Arbeiten im Bereich der Themenextraktion



- Problem: Ausgelegt auf Dokumente

Ansatz

1. Auflösung von Mehrdeutigkeiten mit [Mih07]
2. Themenextraktion mit [HHKG13] und [CMM09]
3. Auswahl passender Ontologien

Auflösung von Mehrdeutigkeiten

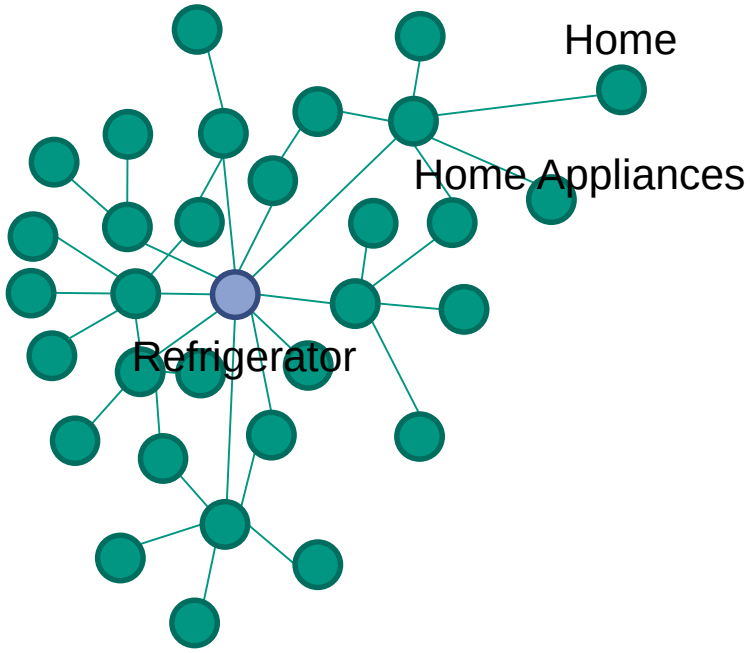
- Ansatz [Mih07] von Mihalcea
  - Auflösung der Mehrdeutigkeiten von Nomen
  - Wikipedia als großer annotierter Korpus

> Cicero describes visiting the tomb of Archimedes, which was surmounted by a sphere and a [[cylinder (geometry)|cylinder]], which Archimedes had requested to be placed on his tomb, representing his mathematical discoveries.

- Erstellen von Trainingsinstanzen
- Training eines Naive-Bayes-Klassifikators
- Anpassungen
  - Log-Sum-Trick gegen arithmetische Unterläufe
  - Gewichtung

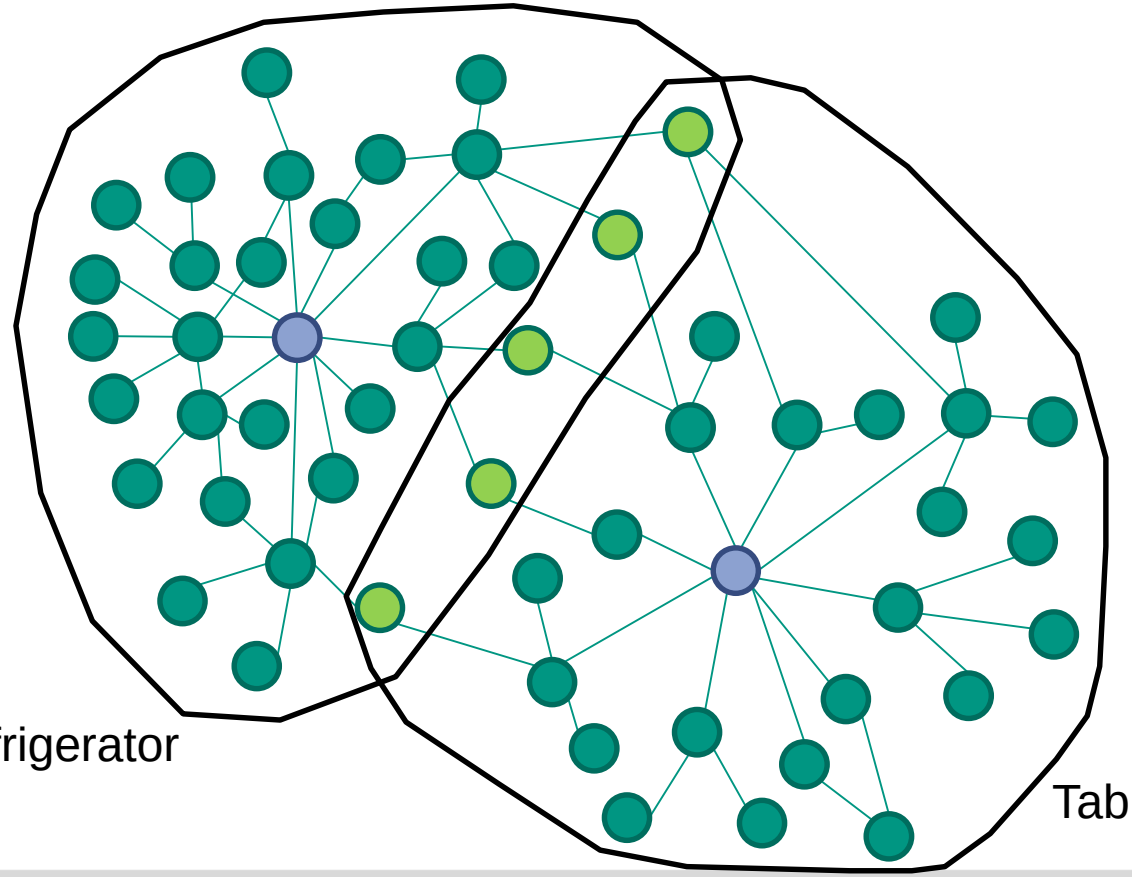
Themenextraktion

- Alle disambiguierten Nomen der Eingabe betrachten
- Jeweils einen Bedeutungsgraph nach [HHKG13] erstellen
  - Verbindungen zu u.a. Kategorie, Hierarchie und Typ



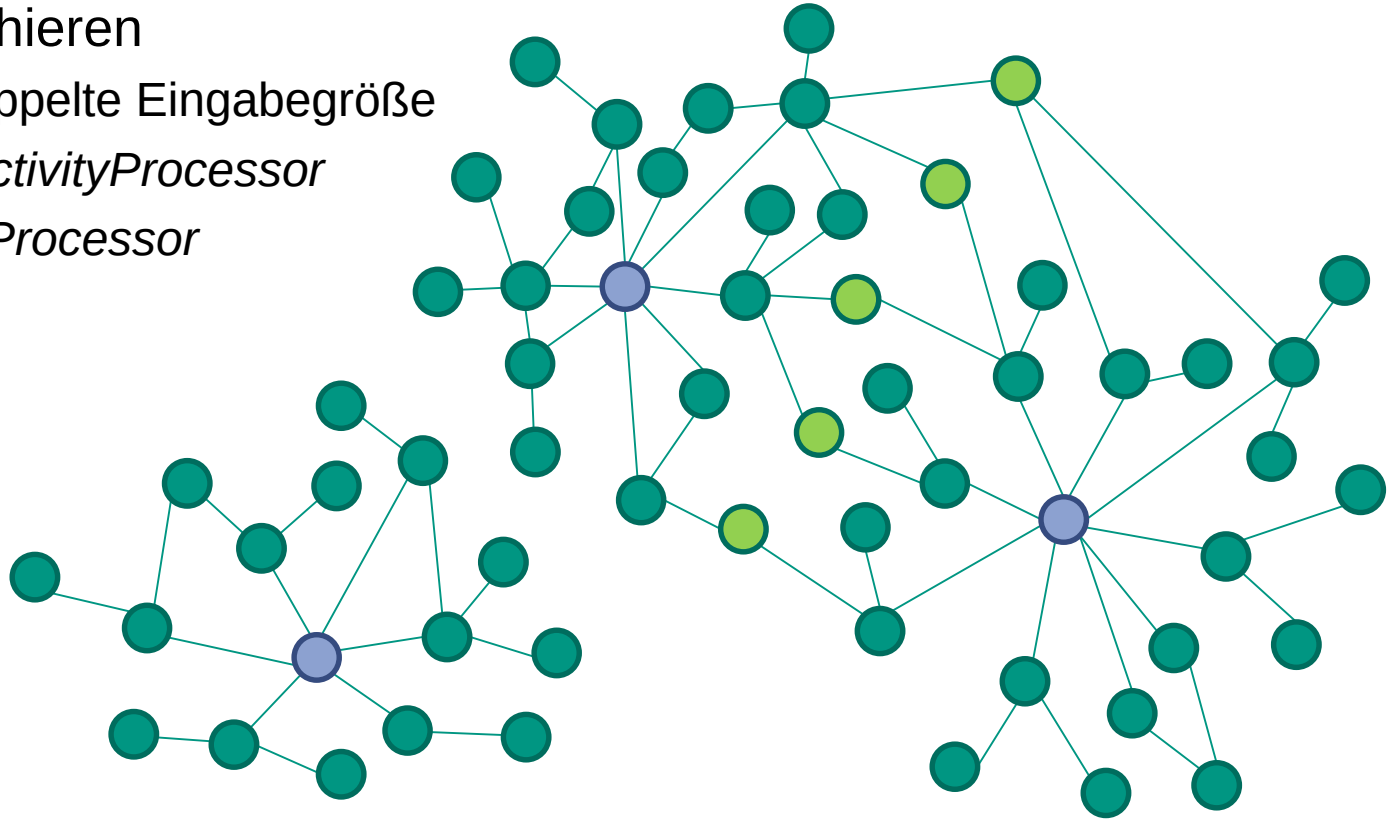
Themenextraktion

- Bedeutungsgraphen zu Themengraph verschmelzen



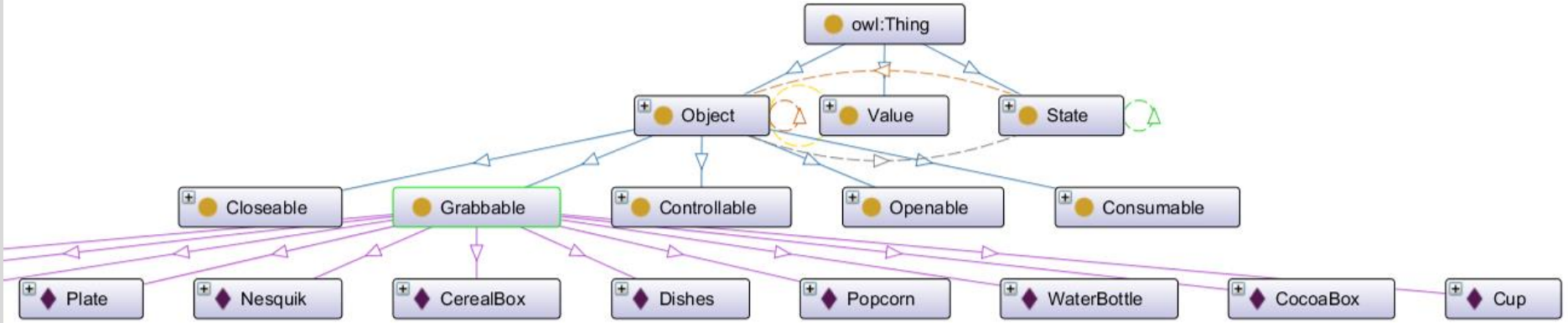
Themenextraktion

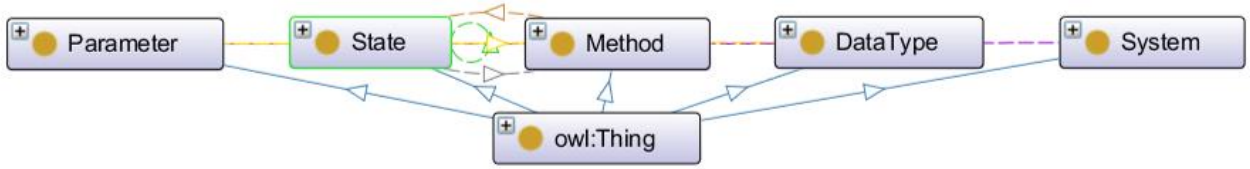
- Themengraph erstellen
- Zentralität bestimmen: Angepasster PageRank-Algorithmus aus [CMM09]
- Konnektivität der Knoten bestimmen
- Themen extrahieren
  - Anzahl: doppelte Eingabegröße
  - *TopConnectivityProcessor*
  - *CombinedProcessor*



Ontologieerstellung

- Trennung zwischen Akteur- und Umgebungsontologien
- Regeln über Struktur und Benennung





Ontologieauswahl

- Themenextraktion auf Ontologien anwenden
- Übereinstimmung der Themen der Anweisungen zu den Themen der Ontologien bestimmen

$$\mathrm{Sim}(t,O) = \begin{cases} 0 & ,t \notin Themengraph \\ \frac{1}{\min\limits_{o \in O}\left(dist(t,o)\right)+1} & ,t \in Themengraph \end{cases}$$

$$C(T,O) = \frac{\sum_{t \in T} Sim(t,O)}{|T|}$$

- Auswahl der passendsten Ontologien mit Hilfe eines Schwellwertes
- Verschmelzung

Evaluation der Auflösung von Mehrdeutigkeiten

- Evaluation mit den erstellten Instanzen
  - 10 Durchläufe mit jeweils 10.000 zufälligen Instanzen
  - Durchschnittliche Genauigkeit: 79,89%
- Evaluation auf dem PARSE-Korpus
  - Anweisungen an Haushaltsroboter ARMAR
  - 8 Szenarien mit insgesamt 169 Aufzeichnungen
    - Bsp.: Anleitung zum Mischen eines Cocktails
  - Genauigkeit von 88,03% (846/961)
- Probleme
  - Korrekte Bedeutung eines Wortes in Wikipedia nicht enthalten
  - Fehlerhafte Annotationen von Wortarten

Evaluation der Themenextraktion

- Umfragen nach Vorbild der Evaluation von Hulpus et al. in [HHKG13]
  - Ein Transkript pro Szenario
    - 8 Szenarien des PARSE-Korpus
    - 3 eigene Szenarien
  - 1. Einstufung: „passend“, „verwandt“, „unverwandt“
  - 2. Einstufung: „zu allgemein“ oder „zu spezifisch“

Evaluation der Themenextraktion

TopConnectivity Processor

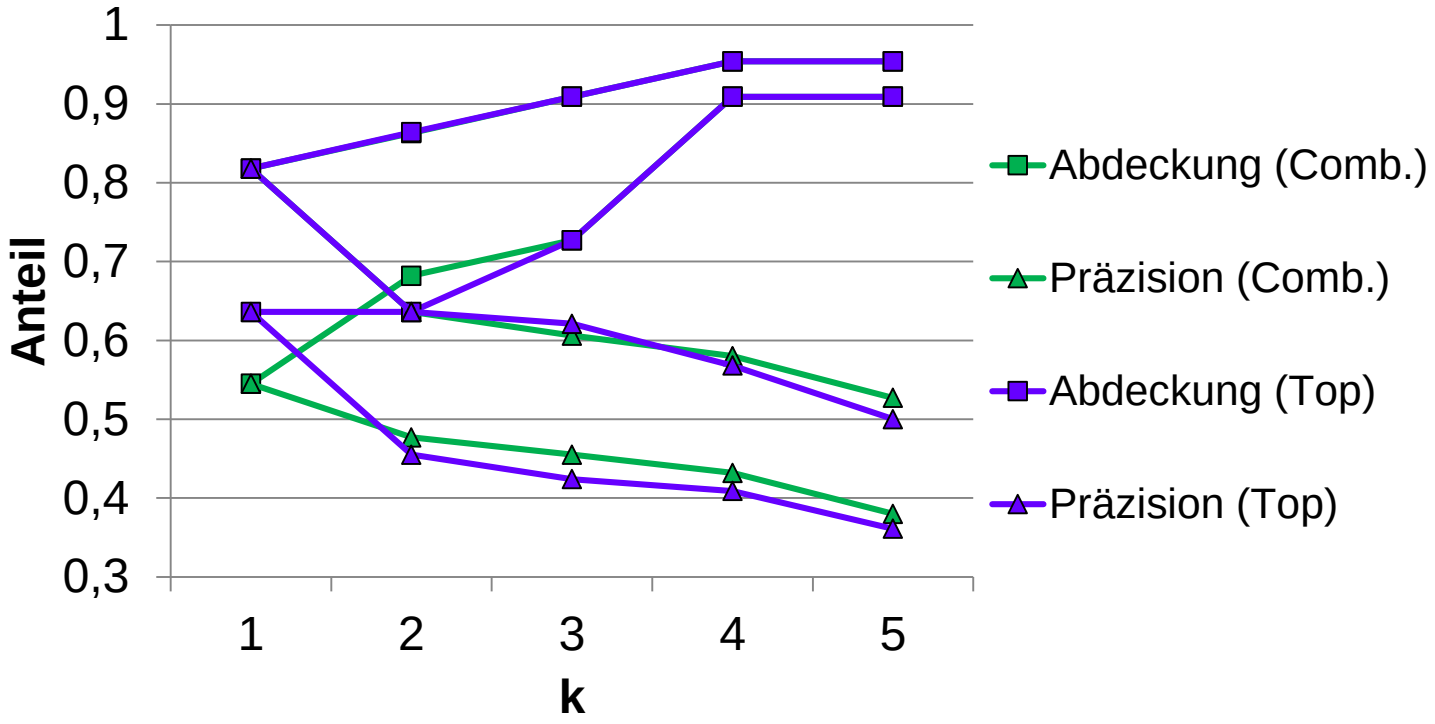
	Passend	Verwandt	Unverwandt
Erstes Thema	53,03%	28,79%	18,18%
Erste 2 Themen	42,42%	28,79%	28,79%
Erste 3 Themen	44,44%	29,80%	25,75%
Erste 4 Themen	42,05%	30,68%	27,27%
Erste 5 Themen	36,54%	31,42%	29,89%
Alle Themen	34,14%	33,26%	32,60%

CombinedProcessor

	Passend	Verwandt	Unverwandt
Erstes Thema	53,03%	28,79%	18,18%
Erste 2 Themen	44,70%	27,27%	28,03%
Erste 3 Themen	44,95%	29,80%	25,25%
Erste 4 Themen	43,18%	29,92%	26,89%
Erste 5 Themen	37,74%	31,72%	29,31%
Alle Themen	36,73%	31,85%	31,42%

Evaluation der Themenextraktion

- Präzision@k = $\frac{\#\text{Treffer in den ersten k Themen}}{k}$
- Abdeckung@k = $\frac{\#\text{Texte mit min. 1 Treffer in den ersten k Themen}}{\#\text{Texte}}$
- „Treffer“: Mehrheit der Befragten schätzt ein Thema passend (oder zu allgemein) ein



Evaluation der Ontologieauswahl (Umgebung)

- 33 Transkripte
  - 16 aus dem PARSE-Korpus
  - 17 selbst erstellte

TE	Ontologieauswahl		Metriken			
#Themen	Schwellwertfaktor	#Themen	Präzision	Ausbeute	F1	F2
2*n	0,10	5	91,89%	89,47%	**90,67%**	**89,94%**
2*n	0,15	5	80,95%	89,47%	85,00%	87,63%
2*n	0,20	5	70,83%	89,47%	79,07%	85,00%
5	0,10	5	85,00%	89,47%	87,18%	88,54%
5	0,15	5	79,07%	89,47%	83,95%	87,18%
5	0,20	5	76,09%	92,11%	83,33%	88,38%
2*n	0,10	10	77,78%	92,11%	84,34%	88,83%
5	0,10	10	79,01%	89,47%	83,95%	87,18%

Zusammenfassung

- Ziele
  - Graphbasierter Ansatz zur Themenextraktion
  - Auswahl von themenspezifischen Ontologien
- Auflösung von Mehrdeutigkeiten nach [Mih07]
  - Wikipedia als Korpus
  - Naive-Bayes-Klassifikator mit Genauigkeit von 88,03%
- Themenextraktion
  - Erstellung Themengraph
  - Extraktion von zentralen, gemeinsamen Knoten
  - Präzision von bis zu 81,8%, Abdeckung von bis zu 95,4%
- Ontologieauswahl
  - Extraktion von Themen
  - Übereinstimmung der Themen
  - F1-Werte von bis zu 90,67%, F2-Werte von bis zu 89,94%

Ausblick

- Themenextraktion für mehrere Themengebiete verbessern
  - Gruppierung innerhalb des Themengraphen, pro Gruppe dann Themen auswählen
- Ansatz für Ontologieauswahl bei Akteur-Ontologien
  - Über Schlüsselbegriffe bzw. Eigennamen
- Themen für Verbesserung anderer Ansätze nutzen
- Mit Themen Informationen aus anderen Quellen gewinnen

Vielen Dank für die Aufmerksamkeit!

Literatur 1/3

- [BNJ03] Blei, David M. ; Ng, Andrew Y. ; Jordan, Michael I.: Latent Dirichlet Allocation. In: Journal of machine Learning research 3 (2003), Nr. Jan, S. 993–1022
- [HCPC13] Hingmire, Swapnil ; Chougule, Sandeep ; Palshikar, Girish K. ; Chakraborti, Sutanu: Document classification by topic labeling. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval ACM, 2013, S. 877–880
- [MCCS09] Magatti, Davide ; Calegari, Silvia ; Ciucci, Davide ; Stella, Fabio: Automatic labeling of topics. In: Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on IEEE, 2009, S. 1227–1232

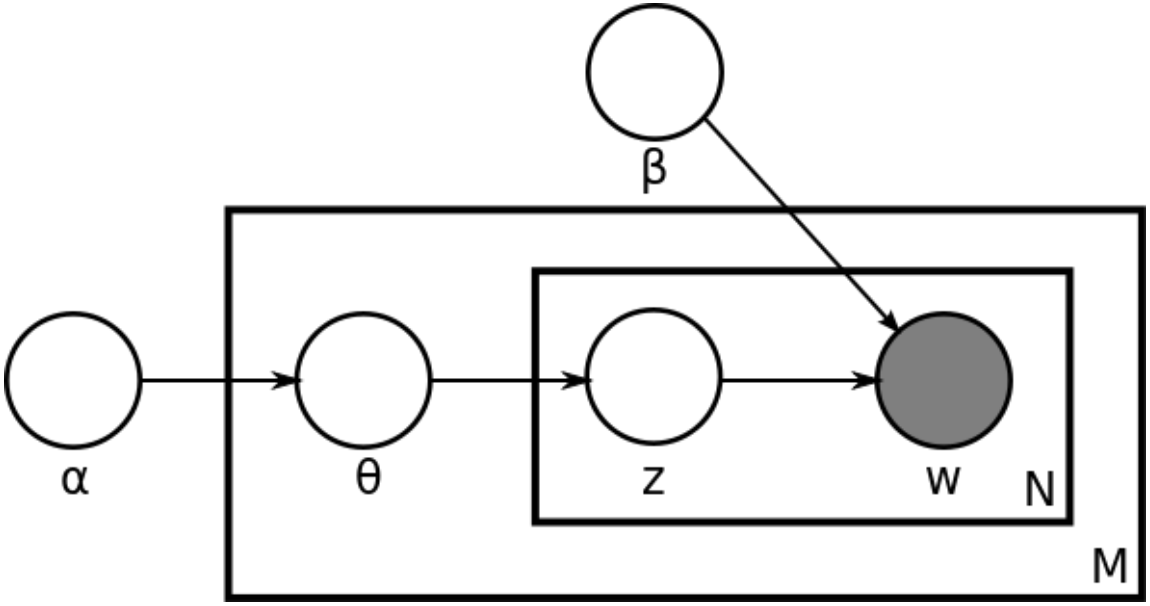
Literatur 2/3

- [HHKG13] Hulpus, Ioana; Hayes, Conor; Karnstedt, Marcel; Greene, Derek: Unsupervised graph-based topic labelling using dbpedia. In: Proceedings of the sixth ACM international conference on Web search and data mining ACM, 2013, S. 465–474
- [CMM09] Coursey, Kino; Mihalcea, Rada; Moen. William: Using encyclopedic knowledge for automatic topic identification. Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009.
- [PBMW99] Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry: The PageRank citation ranking: Bringing order to the web. / Stanford InfoLab.1999.
- [Mih07] Mihalcea, Rada: Using Wikipedia for Automatic Word Sense Disambiguation. In: HLT-NAACL, 2007, S. 196–203

Literatur 3/3

- [LK77] Landis , J R. ; Koch , Gary G.: The measurement of observer agreement for categorical data. In: biometrics (1977), S. 159–174

Latent Dirichlet Allocation (LDA)



N - Wörter im Dokument
M - Dokumente
α - Parameter über Verteilung der Themen pro Dokument
β - Parameter über Verteilung der Wörter pro Thema
θ - Themenverteilung
z - Thema des n‘ten Wortes in Dokument m
w - Wort

PageRank [PBMW99]

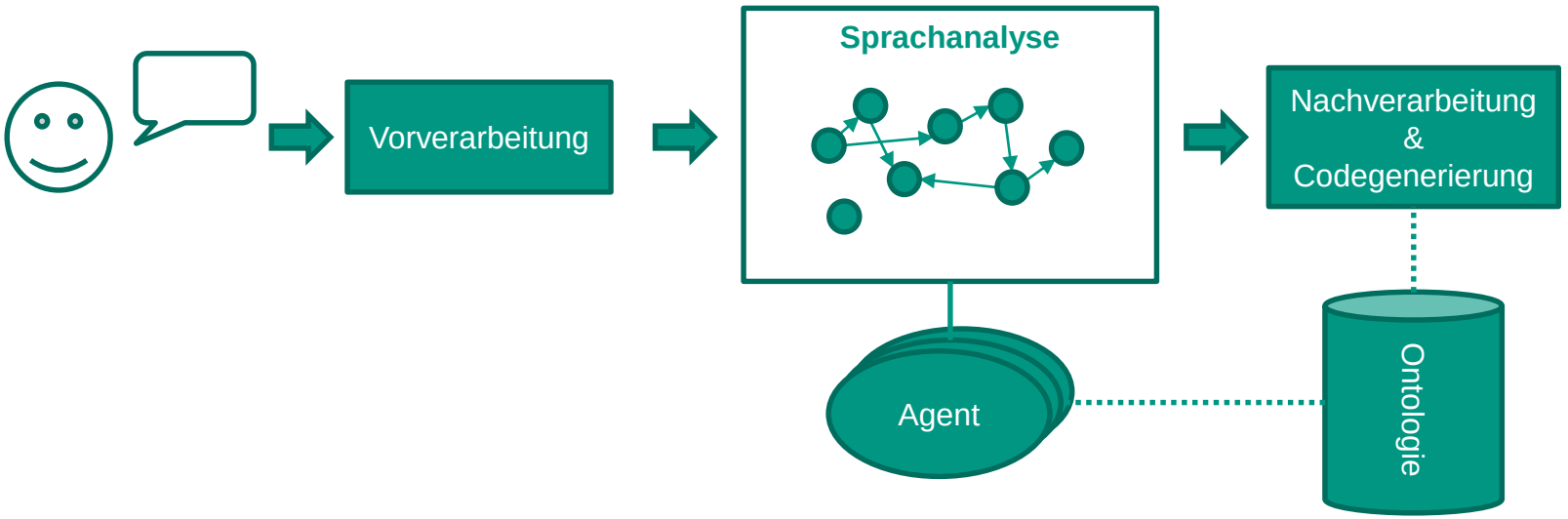
$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_i)|} * S(V_j)$$

Biased PageRank

$$S(V_i) = (1 - d) * Bias(V_i) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_i)|} * S(V_j)$$

Programmieren in natürlicher Sprache

- Projekt PARSE (Programming ARchitecture for Spoken Explanations)
- Aus bekannten Handlungen dem Zielsystem neue Handlungen beibringen



Auflösung von Mehrdeutigkeiten

- Wikipedia-Artikel beschaffen und vorverarbeiten
- Trainingsinstanzen erstellen
  - Keine Eigennamen
  - 5.188.470 Instanzen, 283.173 Klassen
- Training des Klassifikators
  - Anpassung des Naive-Bayes-Klassifikators von Weka
    - Speichereffizienz
    - Logarithmische Werte gegen arithmetische Unterläufe
    - Gewichtung

Naive-Bayes-Klassifikator

$$y = \underset{k \in \{1,\ldots,|C|)}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Log-Sum-Trick:

$$y = \underset{k \in \{1,\ldots,|C|)}{\operatorname{argmax}} \; \log\left(p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \right)$$

$$y = \underset{k \in \{1,\ldots,|C|\}}{\operatorname{argmax}} \left(\log(p(C_k) + \sum_{i=1}^{n} p(x_i|C_k) \right)$$

Genutzte Verbindungen in DBpedia

Bezeichner	Bedeutung
dcterms:subject	Verbindung eines Konzepts zur dazugehörigen Wikipedia-Kategorie
skos:broader	Hierarchie zwischen Kategorien, allgemeiner
skos:broaderOf	Umgekehrte Verbindung von skos:broader
skos:narrower	Hierarchie zwischen Kategorien, spezifischer
purlg:hypernym	Hyperonymie zwischen Konzepten
purlg:meronym	Teil-Ganzes-Beziehungen zwischen Konzepten
purlg:synonym	Synonymie zwischen Konzepten
rdfs:type	Verbindung zu DBpedia-Ontologie-Entitäten
rdfs:subClassOf	Unterklasse innerhalb der DBpedia-Ontologie
rdfs:seeAlso	Ähnliche, weiterführende Konzepte

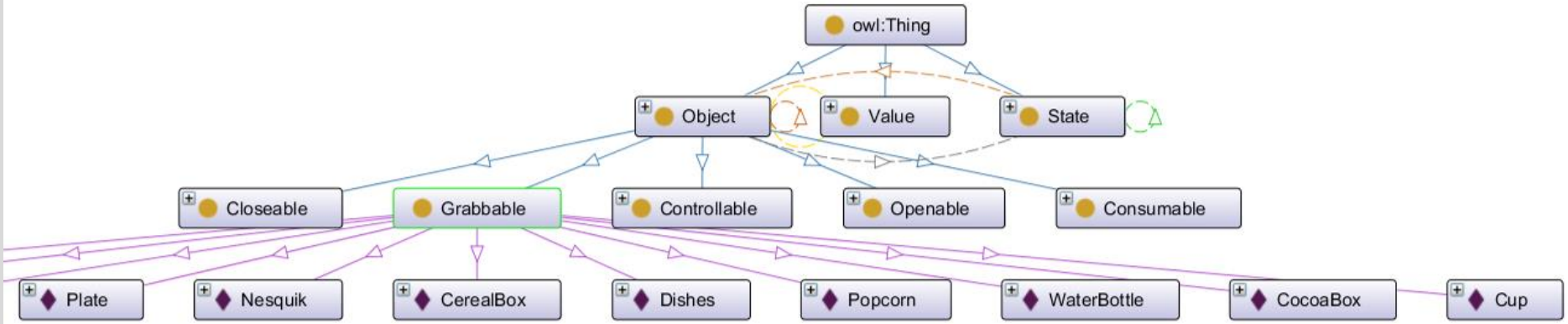
Ontologieerstellung

- 12 Ontologien erstellt
  - 4 Akteur-Ontologien
  - 8 Umgebungsontologien

Typ	Name
Akteur	Household Robot
Akteur	Virtual Assistant
Akteur	Drone
Akteur	Lego Mindstorm

Typ	Name
Umgebung	Kitchen
Umgebung	Bar
Umgebung	Garden
Umgebung	Bedroom
Umgebung	Children‘s Room
Umgebung	Music
Umgebung	Heating
Umgebung	Laundry

Ontologie Kitchen



Auswahl passendste Ontologien

- Schwellwert mit Schwellwertfaktor berechnen

  Beispiel:

  Beste Übereinstimmung 0,40

  Schwellwertfaktor 0,10

  → Schwellwert 0,36

Ontologieverschmelzung

- Ausgewählte Ontologien verschmelzen
- Einfacher Ansatz: Abbildung gleichlautender Entitäten aufeinander
- Zusätzliche Eigenschaft, die Datentypen mit Objekten verbindet
  - Gleichlautende Datentypen mit Instanzen der entsprechenden Objektklasse

PARSE-Korpus

- Anweisungen von Menschen an Haushaltsroboter ARMAR
- Acht verschiedenen Szenarien

Szenario	Kurzbeschreibung
1	Popcorn soll vom Tisch geholt werden
2	Becher vom Tisch soll in Spülmaschine geräumt werden
3	Orangensaft soll aus dem Kühlschrank geholt werden
4	Geschirr soll in Spülmaschine geräumt werden, falls es schmutzig ist
5	Cocktail *Screwdriver* machen, falls möglich mit frischen Orangen
6	Wasser in Tasse füllen und rote Tassen aus Spülmaschine ausräumen
7	Teller säubern, dann Fertiggericht auf dem Teller in Mikrowelle zubereiten
8	Wäsche aus der Waschmaschine nehmen und in den Trockner packen

Evaluation Themenextraktion: Umfrage

- Inter-Annotator-Übereinstimmung (κ)
  - 0,243 und 0,267 (Hulpus: 0,27)
  - „angemessene Übereinstimmung“ („fair agreement“) [LK77]

Evaluation der Ontologieauswahl

- 33 Transkripte
  - 16 aus dem PARSE-Korpus
  - 17 selbst erstellte

Szenario	Synthetisch	#Texte	Domäne
E1	Ja	2	Garten
E2	Ja	2	Kinderzimmer
E3	Ja	2	Heizung, Musik
E4	Ja	2	Garten
E5	Ja	2	Bar
E6	Ja	2	Schlafzimmer
E7	Ja	2	Musik
E8	Ja	2	Musik, Bar
E9	Ja	1	Küche, Garten