Mining concept-to-concept relations

Ruprecht-Karls-Universität Heidelberg
Seminar für Computerlinguistik
Software Projekt WS 2010/2011
Anna Kulpinska, Tomai Grigoriu, Xiaolin Bao

1. Einleitung

	< Arg1, NE 1	Relation,	Arg2> NE 2
Start	Germany	capital	Berlin
Ziel	< Arg1, Class	Relation,	Arg2> Class
	Country	capital	City

DBpedia

```
<http://dbpedia.org/resource/Germany>
<http://dbpedia.org/ontology/capital>
<http://dbpedia.org/resource/Berlin>

<http://dbpedia.org/resource/Germany>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Country>
```

▪Ziele

•Konzepte, die in einer Relation zu einander stehen, zu erzeugen und zu generalisieren.
•Mapping von Wikipedia zu Wordnet

▪Relationen

- **<capital>** *Berlin is the capital of Germany.*
- **<author>** *Goethe is the author of „Faust“*
- **<product>** *Microsoft Office is a product of Microsoft.*
- **<spouse>** *M.Monroe was a spouse of Joe DiMaggio.*
- **<currency>** *Yuan is the national currency of China.*

▪ Computerlinguistische Themen

- Information Extraction (IE)
- Syntaktische Disambiguierung
- Klassenerkennung

▪ Anwendungsbereiche

- Informationsgewinnung
- Informationsextraktion
- Frage-Antwort Systeme
- automatisierte Erstellung von Ontologien

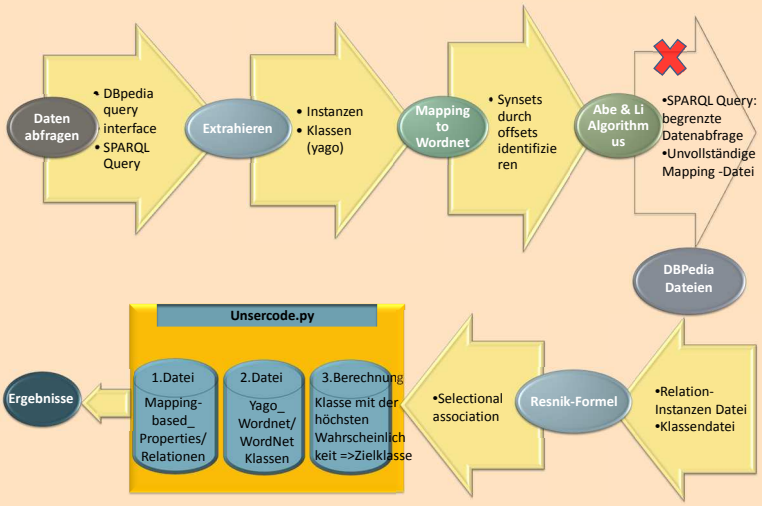
2. Ressourcen

DBpedia	DBpedia ist ein Gemeinschaftsprojekt der Universität Leipzig, der Freien Universität Berlin und OpenLink Software, um strukturierte Informationen aus Wikipedia zu extrahieren und Web-Anwendungen zugänglich zu machen. Im November 2010 enthielt DBpedia 3,5 Millionen Datensätze. Als Standard für die Daten wird das Ressource Description Framework (RDF) benutzt.
WordNet A lexical database for English	Das WordNet ist ein seit 1985 am *Cognitive Science Laboratory* der Princeton University entwickelter Wortschatz der englischen Sprache. WordNet besteht aus einer Datenbank, die semantische und lexikalische Beziehungen zwischen den Wörtern enthält.
Yago/Yago2	YAGO=YAGO2 ist eine riesige semantische Wissensbasis (10 Milionen Einheiten), die aus Wikipedia, WordNet und GeoNames abgeleitet ist.
python	Python ist eine universelle höhere Programmiersprache.
SPARQL	**SPARQL** ist eine graph-basierte Anfragesprache für RDF. *PREFIX dbo: http://dbpedia.org/ontology/* *SELECT ?country ?capital WHERE { ?country dbpedia-owl:capital ?capital. ?country rdf:type yago:Country108544813 . OPTIONAL{?country dbpedia owl:dissolutionYear ?dissolutionYear} FILTER (!bound(?dissolutionYear)) } ORDER BY ?country*

4. Evaluation

	Input	Output	Bewertung
DPedia Ontology	The_Sorrows_of_Young_Werther author Johann_Wolfgang_von_Goethe	→do:Book author do:Person	**1**
	France currency Euro	→owl#Thing currency do:Currency	**1/0**
	France capital Paris	→ do:Country capital do:Place	**1**
	Albert_Einstein birthPlace Ulm	→ do:Person birthPlace do:Place	**1**
WordNet	The_Sorrows_of_Young_Werther author Johann_Wolfgang_von_Goethe	→ not found.	**0**
	France currency Euro	→not found	**0**
	France capital Paris	→ not found	**0**
	Albert_Einstein birthPlace Ulm	→ wn:synset-scientist-noun-1 birthPlace wn:synset-city-noun-3	**1**
Yago	The_Sorrows_of_Young_Werther author Johann_Wolfgang_von_Goethe	→ not found.	**0**
	France currency Euro	→ not found.	**0**
	France capital Paris	→ not found.	**0**
	Albert_Einstein birthPlace Ulm	→ not found.	**0**

3. Arbeitsablauf



▪ Selektive Assoziativität nach Phillip Resnik

$$A(pi,c)=Pr(c/pi)*\log\frac{Pr(c/pi)}{Pr(c)}$$

pi - Prädikat/Relation
c - eine Klasse
Pr(c/pi) - die Wahrscheinlichkeit für das Auftreten der Klasse c bei einem Satz (Zeile in der Datei) mit Relation pi;
Pr(c) – die Wahrscheinlichkeit für Auftreten der Klasse c in irgendeinem Satz unabhängig von der Relation

▪DBPedia Dateien

mappingbased_properties_en.nt – die Datei enthält alle Relationen zwischen Instanzen/Ressourcen von Dbpedia

<http://dbpedia.org/resource/Amsterdam>
<http://dbpedia.org/ontology/country>
<http://dbpedia.org/resource/Netherlands> .

yago_links.nt – die Datei enthält Instanzen/Ressourcen von DBpedia und Wikipedia und dazugehörigen Yago-Klassen

<http://dbpedia.org/resource/Alexander_Mackenzie> rdf:type
<http://dbpedia.org/class/yago/Person100007846> .

wordnet_links.nt – die Datei enthält Ressourcen von DBpedia, deren WordNet 3.0-Klassen zugeordnet sind

<http://dbpedia.org/resource/The_Rootsman> <http://dbpedia.org/property/wordnet_type>
<http://www.w3.org/2006/03/wn/wn20/instances/synset-musician-noun-1>.

instance_types_en.nt – die Datei enthält Ressourcen von DBpedia und dazugehörige Klassen aus DBpedia Ontology

<http://dbpedia.org/resource/Algeria>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> http://dbpedia.org/ontology/Country .

▪Ergebnisse

PREDICATE: author subject -> object	Marilyn_Monroe	Arthur_Miller	Bewertung
(1.1193971924369397,	(ya:Person100007846',	ya:Person100007846'))	**1 (gesuchte Klassen!)**
(0.94900151849862,	(ya:Actor109765278',	'ya:Person100007846'))	**1**
(0.6565943677508124,	(ya:Person100007846',	ya:Actor109765278'))	**0**
(0.5886637261938659,	(ya:Actor109765278',	ya:Actor109765278'))	**0**
(0.5291395214417519,	(ya:Sovereign110628644',	'ya:Person100007846'))	**0**

5. Referenzen

•N. Abe, H. Li. Learning Word Association Norms Using Tree Cut Pair Models. Theory NEC Laboratory, RWCP∗c/o C & C Research Laboratories, NEC, ICML 1996.
•Philip Resnik. Selectional constraints: an information-theoretic model and its computational realization. Cognition 61 (1996) 127-159
•Sergey Brin. Extracting Patterns and Relations from the World Wide Web. Computer Science Department Stanford University.1999
•M. Banko, M. J Cafarella, S. Soderland, M. Broadhea, O.Etzioni. Open Information Extraction from theWeb. Department of Computer Science and Engineering University of Washington Box 352350 Seattle, WA 98195, USA, In IJCAI (2007), pp. 2670-2676.
•S. P. Ponzetto, R. Navigli. Large-Scale Taxonomy Mapping for Restructuring and IntegratingWikipedia. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), 2009, pp. 2083-2088.
• http://www.python.org/
•http://dbpedia.org
•http://dbpedia.org/sparql
•http://dbpedia.org/snorql/
•http://de.wikipedia.org
•http://www.cl.uni-heidelberg.de/~ponzetto/wikitax2wn/