Extraktion von Instanzen semantischer Relationen

Softwareprojekt von Amol Phadke, Jutta Pieper, Joseline Saamago, Robert Schumann
Institut für Computerlinguistik, Universität Heidelberg
WiSe 2012/13

1. ZIEL

- Finden von Subjekten, die in einer bestimmten semantischen Relation zueinander stehen *(wie z.B. Hyponymie oder Meronymie)*, dies soll für jede beliebige Relation funktionieren
- mögliche Anwendungsgebiete: *Word Sense Disambiguation, Ontologien, Information Retrieval, Semantische Annotation*

2. METHODE

BOOTSTRAPPING
- zirkulärer Prozess bei dem aus einer geringen Menge an hochwertigen Daten neue Daten gewonnen werden
- in jedem Iterationsschritt werden Muster extrahiert, mithilfe deren neue Daten gewonnen werden
- die neuen Daten dienen wiederum als Eingabe des nächsten Iterationschrittes

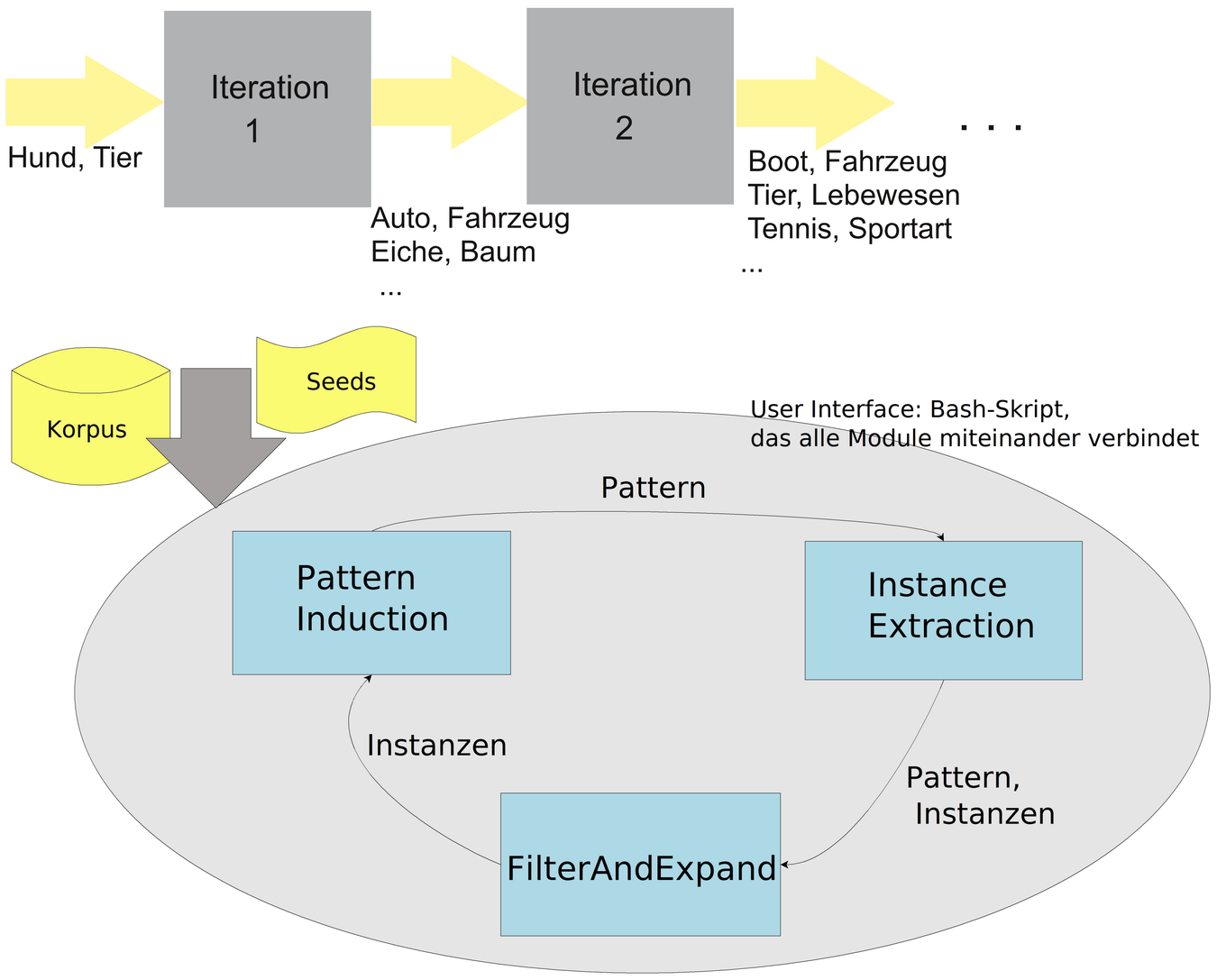
3. PROBLEME BEIM BOOTSTRAPPING

PROBLEME:
- nicht für kleine Korpora geeignet
- genaue Pattern ergeben korrekte aber wenige Instanzen
- ungenaue Pattern ergeben viele aber auch inkorrekte Instanzen

LÖSUNGSANSATZ:
- Verwendung des Webs als zusatzliches Korpus
- Berechnung von Verlässlichkeits- und Konfidenzmaß der Instanzen

4. ESPRESSO

- Implementierung des Espresso-Algorithmus von Pantel & Pennacchiotti
- **Bootstrapping Algorithmus** mit **Filterung** der Instanzen anhand des Verlässlichkeitsmaßes und Weiterverarbeitung zur Extraktion von neuen Instanzen
- **Verlässlichkeitsmaß:** Maß darüber, wie stark ein Pattern mit zwei Instanzen assoziiert ist
- Extraktion der Instanzen aus einem **Korpus** bzw. aus dem **WWW**



5. SYSTEMARCHITEKTUR

- **Eingabe**: Ein **Korpus** und eine **manuell erstellte** kleine Menge an Wortpaaren, die sogenannten **Seeds**, die in einer semantischen Relation zueinander stehen
- **Pattern Induction**: **Extraktion** der Pattern, d.h. **Muster**, die zwei Instanzen verbinden, aus dem Korpus **mithilfe der Seeds**
  Beispiel Pattern:
  Dieser/PDAT/dies TR_X/TR/TR ist/VAFIN/sein TR_Y/TR/TR TR/TR/TR und/KON/und.
  * TR steht für eine Nominalphrase
- **Instance Extraction**: Extraktion **neuer Instanzen**, die in einer semantischen Relation zueinander stehen, aus dem Korpus **mithilfe** der im Modul Pattern Induction **extrahierten Muster**
- **FilterAndExpand**: **Filterung** der im Modul Instance Extraction extrahierten Instanzen **mithilfe eines Vertrauensmaßes** sowie Gewinnung **neuer Instanzen** bei Bedarf aus dem **WWW**
- Weitergabe der neuen Instanzen zusammen mit einem berechneten Velässlichkeitsmaß an das Modul Pattern Induction und Wiederholung des Prozesses zur Gewinnung neuer Instanzen von beliebigen semantischen Relationen

Relation : is a					Relation : part of		
# Tokens	**Iteration**	**#Instances**	**Precision Annot. 1**	**Precision Annot. 2**	**Instances**	**Precision Annot. 1**	**Precision Annot. 2**
87 000	1	1	100.00%	100.00%	0	-	-
	2	1	100.00%	100.00%	0	-	-
	3	1	100.00%	100.00%	0	-	-
884 800	1	11	27.27%	54.00%	15	80.00%	80%
	2	12	25.00%	58.33%	15	80.00%	80%
	3	12	25.00%	58.33%	15	80.00%	80%
1 773 847	1	17	17.64%	35.29%	76	19.74%	28.94%
	2	30	16.66%	20.00%	75	12.00%	16.00%
	3	37	20,27%	21.62%	122	17.21%	13.11%

6. EVALUATION

- Test des Programms auf verschieden große Ausschnitte aus DeWac (hierbei wurde die Verlässlichkeitsschwelle nicht gesetzt, d.h. normales Bootstrapping wurde getestet)
- zwei Teilnehmer (Annot. 1 + Annot. 2) haben die Ergebnisse manuell mit folgenden Werten versehen:
  1 - die Instanz gehört zu der Relation
  0 - sie gehört nicht dazu
  aus den Werten wurde die Precision berechnet
- die Tabelle zeigt die Ergebnisse:
  - die Anzahl der gefundenen Instanzen verdeutlicht, das Bootstrapping nicht gut für kleine Korpora geeignet ist

7. AUSBLICK

- Optimierung des Verfahrens, indem der Einfluss verschiedener Parameterwerte auf die Ergebnisse studiert wird (und die Default-Werte der Parameter angepasst werden)
- Berechnung von relative recall durch Vergleich der Ergebnisse verschiedener "Systeme", also Durchläufe mit anderen Parametern
- Optimierung der Pattern durch Experimente mit unterschiedlich großen Kontextfenstern und/oder durch alternative Verwendung von POS-Tags statt Lemmata

8. RESSOURCEN

- DeWaC-Korpus: erstellt durch Web-Crawling für Webseiten der Domäne .de
- faroo-API: eine "freie" API Alternative zu komerziellen Suchmaschinen

LITERATUR

Pantel & Pennacchiotti (2006):
Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations

Pantel & Ravichandran (2004):
Automatically Labeling Semantic Classes

Manning & Schütze (1999):
Foundations of Statistical Natural Language Processing