Ruprecht-Karls-Universität Heidelberg
Institut für Computerlinguistik
Softwareprojekt WS 2012/13
Andreas Bischoff, Andreas Gölz

Bootstrappingzyklus zur Erlernung semantischer Klassen von Nomina

Motivation

Unterteilung von Wörtern in semantische Kategorien anhand von Kontextmerkmalen.

- Es sollen keine Regeln geschrieben werden.
- Eine Anzahl an Instanzen - bis zu 100 Wörter - ist gegeben (annotierte "Seed"-Nomen).

Lösung: Bootstrapping

Wiederhole folgende Schritte:

Lerne Features anhand von Instanzen
Lerne Instanzen anhand der gefundenen Features

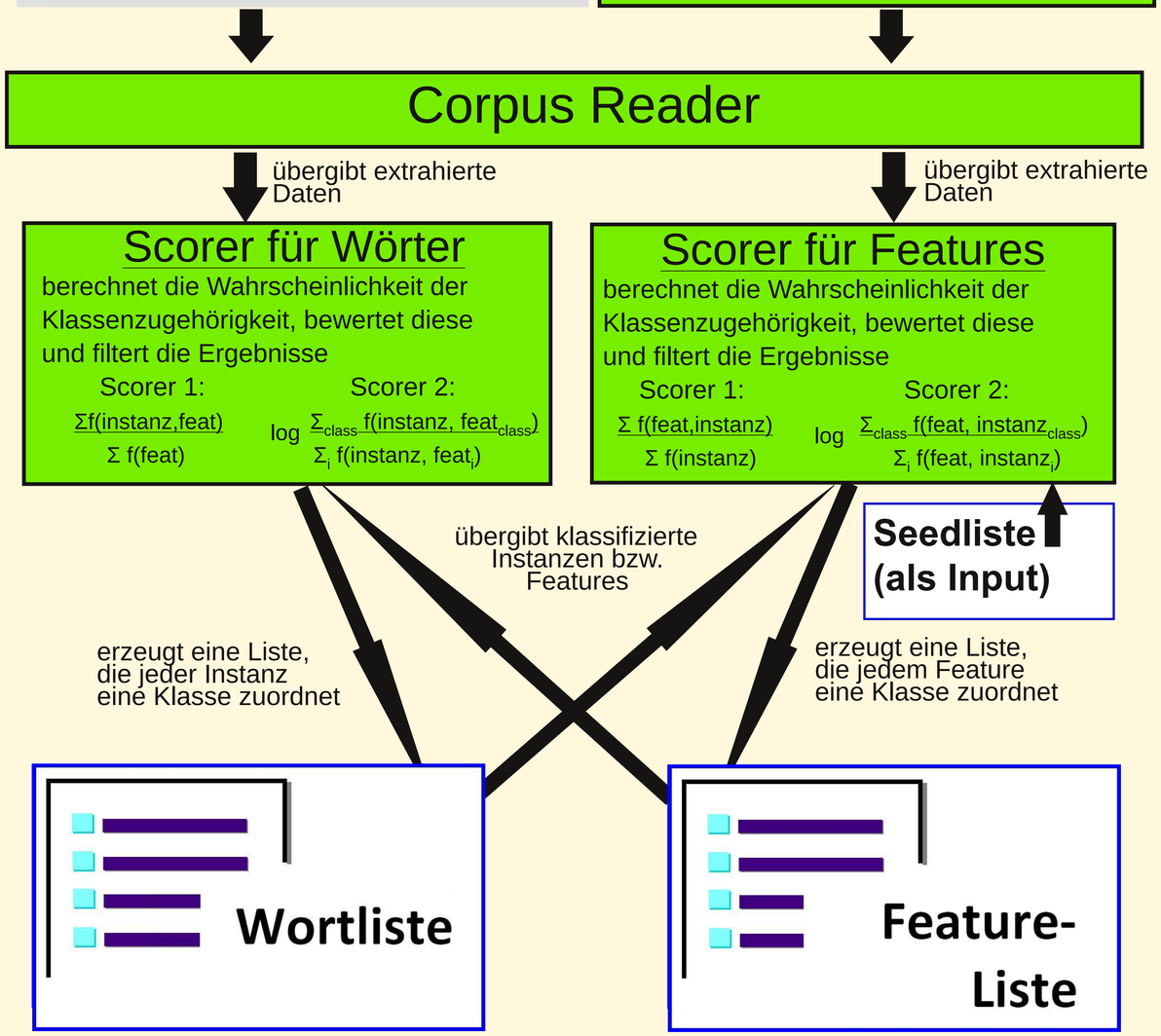
Bei der Evaluierung liegt das Augenmerk auf folgender Klassifizierung:

Objektnomen:
konkret ("Buch") / abstrakt ("Freiheit")
Ereignisnomen:
Ereignisse ("Weihe") bzw. Zustände ("Heiterkeit")
ambige Fälle:
"Absperrung": Kann ein Zaun sein oder eine Arbeit

Die Nomina sind im Experiment die Instanzen, deren Kontexte die Merkmale.

Beispiele (Gorzitze und Padó : 2012)

Verb-Merkmale (Ereignisse) anzetteln-OBJ, verstreichen-SUB, ableisten-OBJ, ...	**Verb-Merkmale (Objekte)** trinken-OBJ, erkranken-SUB, errichten-OBJ, ...
Nominalmerkmale (Ereignisse) Ableistung, Beendigung, Vorabend, ...	**Nominalmerkmale (Objekte)** Seele, Osten, Beziehung, ...
Adjektiv-Merkmale (Ereignisse) erkennungsdienstlich, mündlich, 30jährig, ...	**Adjektiv-Merkmale (Objekte)** mittelständisch, rund, gelb, ...



Resultate

Grundlage für den Goldstandard: "sdewac-nomen" - 87892 Nomina, Korpusfrequenz angegeben.

Aus diesem wurden dreimal 100 Wörter extrahiert ("Common 100", "Common 500", "Common 1000") und per Hand nach ihrer Kategorie (Objekt, Ereignis, ambig) annotiert.

Output des Bootstrappingzyklus, durch zwei Scorer bewertet, wird mit dem annotierten GS verglichen.

Ergebnisse:

Die beiden Scorer zeigen keine signifikanten Unterschiede. Es zeigt sich, dass die niedrigfrequenten evaluierten Wörter die besten Durchschnittswerte erzielen.

Problem: Schwierige Klassifizierung ambiger Wörter!

	Event	Objekt	Ambig	Ø
Prec.	23.81%	27.01%	53.85%	28.67%
Rec.	23.81%	67.86%	9.15%	28.67%

Tab 1.: Ergebnis von Scorer 1 (berechnet wurde der Durchschnitt der drei Frequenzstufen)

Anwendungen des Verfahrens:

- Informationsextraktion (Ereignisse prominent)
- Fragenbeantwortung (je nach Kategorie unterschiedlicher Fragentyp)
- Messung der Lesegeschwindigkeit

Literatur:

S. Gorzitze, S. Pado (2012). *Corpus-based Acquisition of German Event- and Object-Denoting Nouns.* In: Proceedings of KONVENS 2012, pp. 259-263. Wien.

M. Thelen, E. Riloff (2002): *A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts.* In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 214-221. Philadelphia.