

Automatische Extrahierung von semantischen Relationen aus einsprachigen Wörterbüchern

Magisterarbeit in Vorbereitung
Anna Björk Nikulásdóttir
Universität Heidelberg
01.06.05

Gliederung des Vortrags

- Ziel und Motivation der Arbeit
- Verwandte Projekte
- Geplanter Aufbau der Arbeit

Ziel der Arbeit

- Möglichkeiten der Extrahierung semantischer Relationen von Substantiven, Verben und Adjektiven aus einem Isländisch-Isländischen Wörterbuch zu untersuchen
- Eine Menge von extrahierten Relationen zu gewinnen
- Analysen der Definitionsmuster des Wörterbuchs bereitstellen

Motivation

- bisher kein “IceNet” vorhanden
- kleine Sprachgemeinschaft (etwa 300.000 Sprecher) ⇒ begrenzte Kapazitäten für manuelles Erstellen eines Wortnetzes
- Die Analyse der Definitionsmuster wird für die Überarbeitung des Wörterbuchs benötigt

Warum Wörterbuch?

- explizite Informationen zu sehr vielen Wörtern der Sprache
- eindeutige Angaben zu grammatischen Eigenschaften eines Lemmas
- relativ zuverlässige Strukturen der Definitionstexte

Probleme

- Printwörterbücher sind nicht für die maschinelle Verarbeitung konzipiert
- Nicht alle Definitionstexte weisen eindeutige Strukturen auf
- Kreisdefinitionen: ein Wort A wird mit einem Wort B erklärt und Wort B wird mit dem Wort A erklärt.
- und wahrscheinlich noch viele mehr ...

Beispiel: “schnell”

schnell <Adj.> **1** *rasch, geschwind, eilig, flink*; Ggs. *langsam(1)*; eine ~e Bedienung; ein ~es Pferd; [...]

Beispiel: “Auto”

Auto <n.; -s, -s; Kurzw. für> *Automobil*; jmd. kann gut, schlecht ~ fahren; das Autofahren mit Kindern kann im Sommer zur Qual werden

Automobil <n.; -s, -e; Kurzw.: Auto> *Personenkraftwagen*

Personenkraftwagen <m.; -s, -; Abk.: Pkw> *Kraftwagen zum Befördern von Personen; Sy Auto, Automobil; Ggs Lastkraftwagen*

Kraftwagen ⇒ **Kraftfahrzeug** ⇒ **Landfahrzeug**

Beispiel: “fahren”

fahren <V. 130> **2** <400(s.)> *sich mit einem Fahrzeug fortbewegen; Ggs. gehen(1); wir wollen lieber ~ (anstatt zu gehen); [...]*

Tools, Ressourcen

- Wörterbuchdatenbank des Verlags, welches das Wörterbuch herausgibt (ca. 105.000 Stichwörter)
- Tagger für das Isländische

Gliederung des Vortrags

- Ziel und Motivation der Arbeit
- **Verwandte Projekte**
- Geplanter Aufbau der Arbeit

Verwandte Projekte

- Automatische Extrahierung von WordNet Relationen
- eXtended WordNet
- Automatische Extrahierung von semantischen Relationen aus einem Baskischen Wörterbuch
- Extrahierung einer Hyperonymiehierarchie aus WDG

Automatische Extrahierung von WordNet Relationen

- Erweiterung des manuell erstellten WordNets
- Definitionstexte aus Enzyklopädiën und Zeitungstexten
- Patterns vs. syntaktisches Parsing

Lexicosyntactic Patterns

Agar is a substance prepared from a mixture of red algae, **such as Gelidium**, for laboratory or industrial use.

a. NP_0 *such as* NP_1 {, NP_2 ..., (*and/or*) NP_i } $i \geq 1$

b. for all NP_i , $i \geq 1$, HYPONYM (NP_i , NP_0)

HYPONYM (**Gelidium**, red algae)

eXtended WordNet

- Genaue semantische Annotierung von open-class Wörtern in WordNet Glossen
- Schwerpunkt auf precision, Ziel ist annähernd 100% precision
- Recall nicht so wichtig

eXtended WordNet

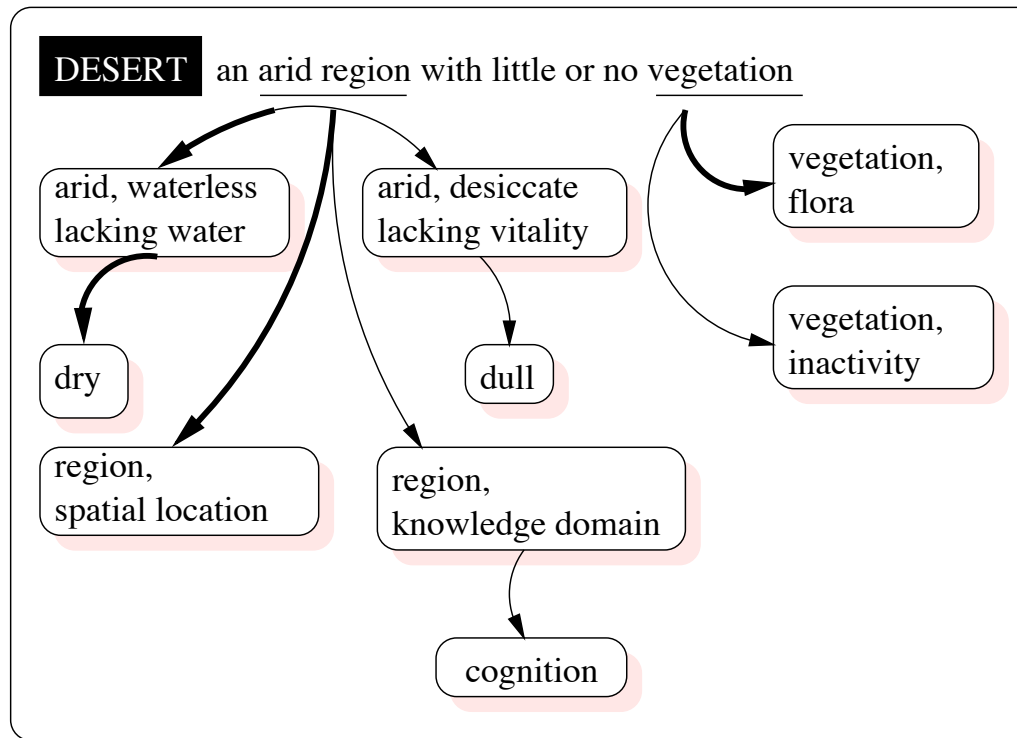
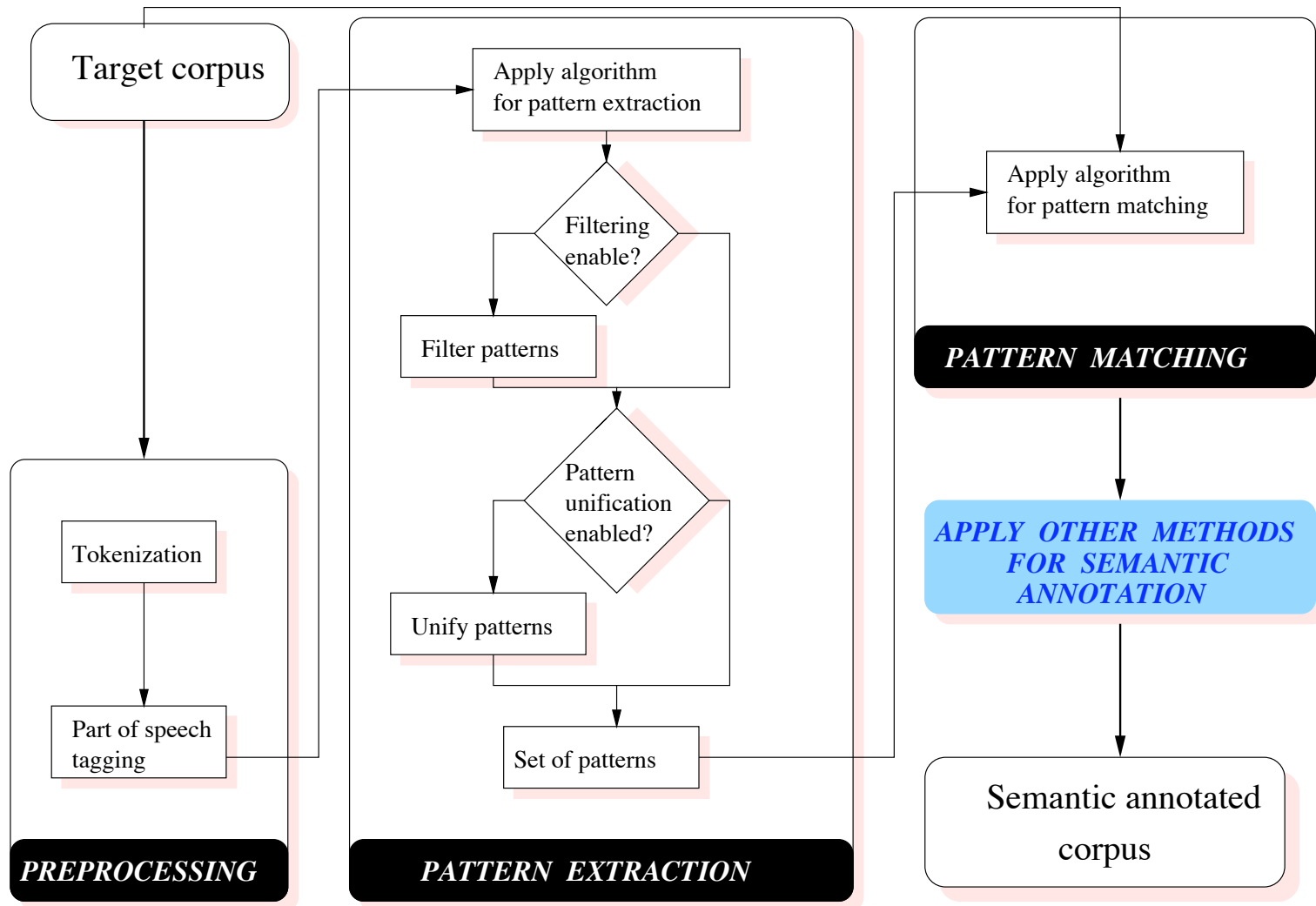


Figure 1: Semantic relations across WordNet concepts

eXtended WordNet



Baskisches Wörterbuch

- Extrahierung von semantischen Relationen aus einem Baskischen Wörterbuch
- Baskisch ist eine agglutinierende Sprache \Rightarrow morphosyntaktische Analyse wichtig
- Syntaxanalyse basiert auf Constraint Grammar
- Regeln basieren auf extrahierten Mustern

Baskisches Wörterbuch

- 42.533 Relationen extrahiert
- 9% der Definitionen, meist Adjektive, ohne Relation
- 2,2% der extrahierten Relationen sind falsch
- Hohe precision und recall

Hyperonymiehierarchie aus dem WDG

- Definitionen getaggt und gehunkt
- Annahmen: Einwortdefinitionen = Synonyme
Mehrwortdefinitionen = Hyperonyme (Kopf)

Hyperonymiehierarchie aus dem WDG

- Precision: Richtig erkannte Köpfe ~70%
Richtig klassifizierte Hyperonym-Hyponym
Paare: ~60%
- Recall: In 94% der Atomardefinitionen wurde
ein Kopf erkannt

Gliederung des Vortrags

- Ziel und Motivation der Arbeit
- Verwandte Projekte
- **Geplanter Aufbau der Arbeit**

Geplanter Aufbau der Arbeit

I. Theorieteil

- semantische Relationen in bestehenden Projekten (WordNet, GermaNet, LeXem...)
- Mikrostrukturen in einsprachigen Wörterbüchern

Geplanter Aufbau der Arbeit

2. Untersuchung

- Mikrostrukturen in dem Isländischen Wörterbuch
- Analyse der Definitionsmuster
- Regeln zur Extrahierung der Relationen
- Ergebnisse und Evaluierung

Bibliographie

- Hearst, M.A. (1998): Automated Discovery of WordNet Relations. in: C. Fellbaum (ed.) WordNet : an electronic lexical database, MIT
- Novischi, A. (2002): Accurate Semantic Annotations via Pattern Matching. <http://xwn.hlt.utdallas.edu/papers.html>
- Agirre, E. et al. (2000): Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar. http://ixa.si.ehu.es/ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000808988
- Geyken, A., Ludwig, R. (2003): Halbautomatische Extraktion einer Hyperonymiehierarchie aus dem Wörterbuch der deutschen Gegenwartssprache. <http://www.bbaw.de/forschung/kollokationen/documents/ExtrHyp.pdf>