

GermaNet und UniNet: Anknüpfen an semantische Netze

Simon Clematide
Institut für Computerlinguistik
Universität Zürich



Überblick



UniNet

- ◆ Inhalt, Struktur, Charakteristik

Vergleichende Mini-Evaluationen der Abdeckung

- ◆ Häufigste und zufällig gewählte Substantive: Germa- vs. UniNet

Verknüpfen von semantischen Netzen

- ◆ Problem und 3 Szenarien
 - ◆ Abgleich von Netz-Versionen
 - ◆ Integration von multilingualen Netzen
 - ◆ Integration von Teil-Netzen
- ◆ Vom ILI zum INI: Interlingual-Index als Paradigma
 - ▶ Anknüpfen durch sparsames, explizites Indizieren

UniNet (1999)



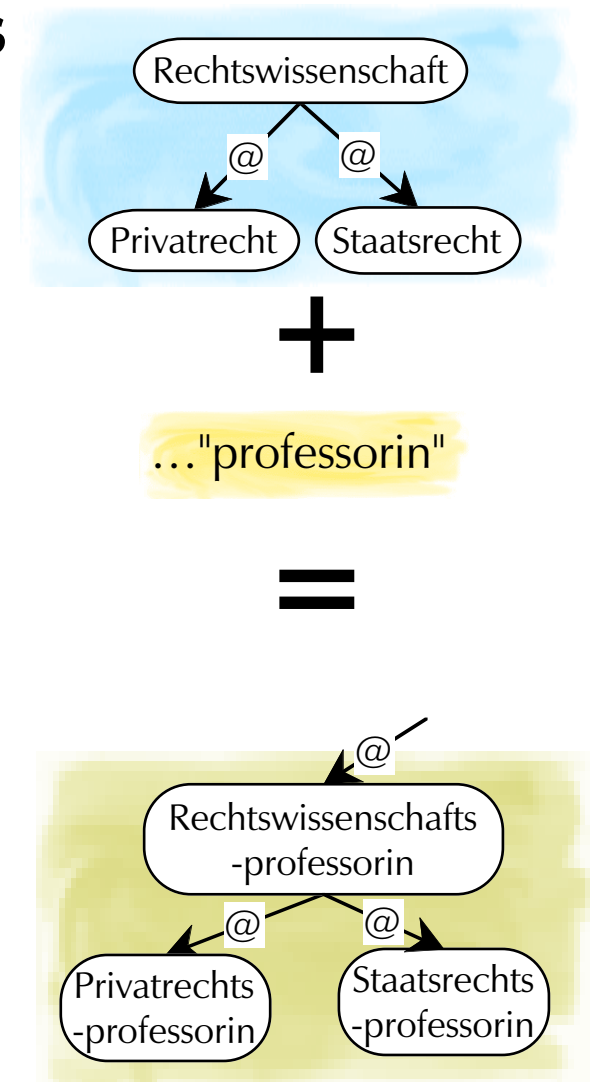
Ein deutsch(-schweizerisch-)es semantisches Netz

- ◆ **Sachgebiet:** Hochschulen
- ◆ **Struktur:** à la GermaNet
 - ◆ Lexikalischer Bestand
 - ◆ 21'901 einteilige Substantive
 - ◆ 73 einteilige Eigennamen
 - ◆ 1'199 mehrteilige Substantive
 - ◆ 494 mehrteilige Eigennamen
 - ◆ Semantischer Bestand
 - ◆ 12'385 Synonymklassen
 - ◆ 6'996 Synonymklassen haben mehr als 1 Oberbegriff
 - ◆ 22'514 Oberbegriffsbeziehungen
 - ◆ 2'648 Teil-Ganzes-Beziehungen
- ◆ **Inhalt:** Weder mit GermaNet 1999 noch 2001 kompatibel...

UniNet: *special effects...*

Automatisches Weiterspinnen des Netzes durch Wortbildungsmuster

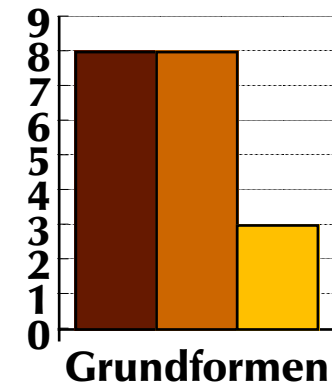
- ◆ Ca. 550 Bezeichnungen für Studienfächer und -gebiete
- ◆ und eine Handvoll Kompositabestandteile
 - ◆ Erstglieder wie "Hauptfach..."
 - ◆ Letztglieder wie "...professorin"
- ◆ ergeben viele neue Synonymklassen mit den entsprechenden semantischen Relationen!
 - ▶ Problem: Übergenerierung im Stil von "Hauptfachkatastrophenmedizinstudentin"...



Mini-Evaluation der Abdeckung I

12 häufigste Wortformen (9 Grundformen)

- ◆ in einem hochschulspezifischen Korpus (350 Sätze)
- ◆ GermaNet 01 und UniNet decken zusammen alle 9 Grundformen ab
 - ◆ "Studierende" fehlt in GN 01, "Paragraph" in UN

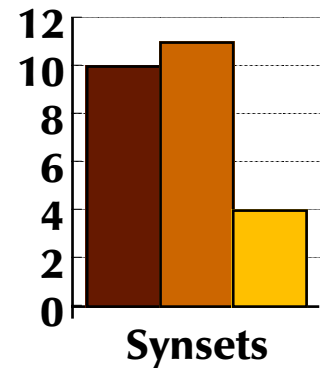


Synonymklassen

- ◆ Zudem: Wegen der Mehrfachvererbung der Oberbegriffsrelation, gilt die Gleichung

$$1 \text{ Synset} = 1 \text{ Lesart}$$

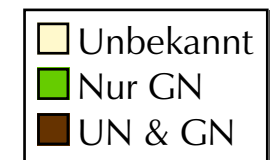
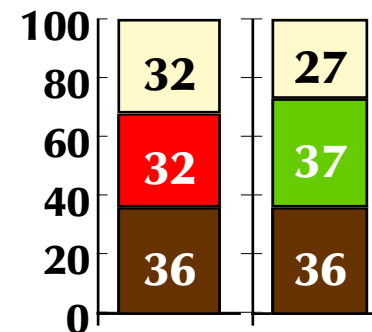
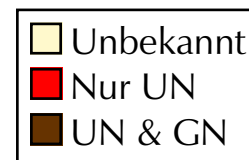
aus WordNet nicht mehr.



Mini-Evaluation der Abdeckung II

100 unterschiedliche zufällig gewählte Substantive

- ◆ 21 Wörter im GN, aber nicht im UN
 - ▶ Aber nur 3 aus engerem Anwendungsgebiet von UN!
- ◆ 16 Wörter im UN, aber nicht im GN
 - ▶ 15 aus engerem Anwendungsgebiet von UN
- ◆ Im Detail recht unterschiedliche Synonymklassen!
- ◆ UN liefert 82 Synonymklassen für 68 Wörter (1,2 W/S)
- ◆ GN liefert 126 Synonymklassen für 73 Wörter (1,7 W/S)



Abdeckungsgrade

Verknüpfen von semantischen Netzen

Utopie: Plug-And-Play mit semantischen Netzen

- ◆ Allgemeines Netz für Grundwortschatz und **ontologische** Grob-Organisation
 - ▶ GermaNet ist z.T. zu spezifisch und ohne Metainformation dazu!
- ◆ Spezialisierte Netze für Anwendungsgebiete mit **terminologischen** Beständen und *named entities*
 - ▶ UniNet ist in der allgemeineren Strukturierung eine Insellösung!

Problem: Integration bestehender Ressourcen

- ◆ Semantische Netze sind "selbstorganisierend" (leicht autistisch)

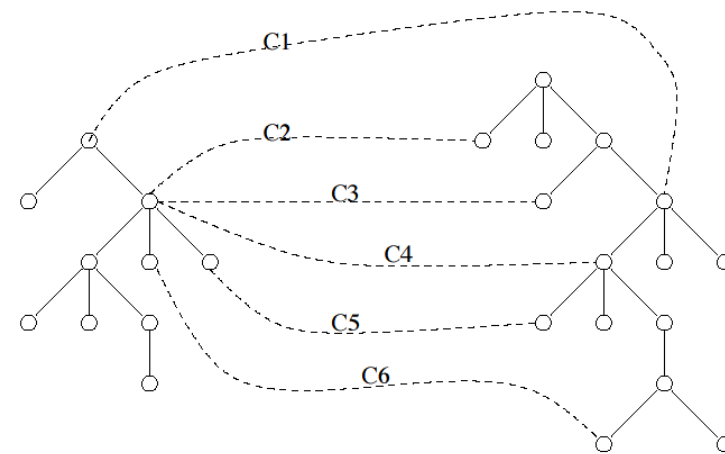
Lösungsansätze

- ◆ Handarbeit vs. Halbautomatisierung vs. Vollautomatisierung
-

Szenario I: Netz-Versionen

Integration von unterschiedlichen Versionen desselben Netzes

- ◆ Beispiel: Mapping von WordNet 1.5 zu WordNet 1.6
 - ▶ Vollautomatisch machbar (*relaxation labelling*) [Daudé et al. 01]
 - ▶ Direktes Verknüpfen von Netz-Elementen
- ◆ Wichtige Eigenschaften
 - ▶ Hohe Strukturähnlichkeit
 - ▶ Grosse lexikalische Überschneidung



Direktes Verknüpfen

Szenario II: Multilinguale Netze

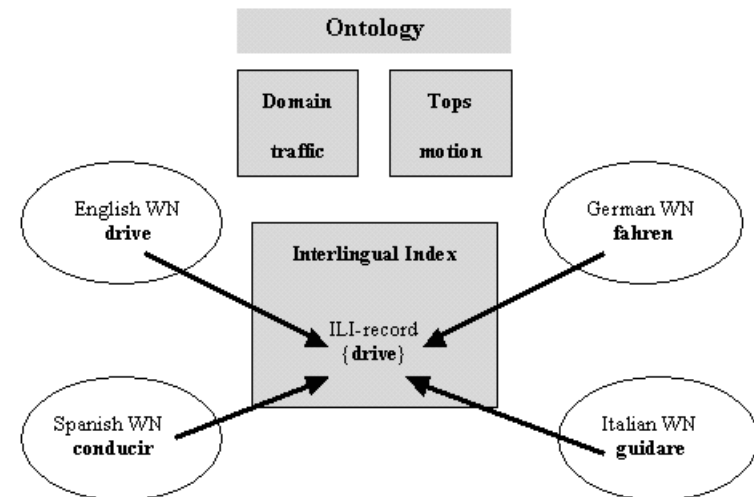
Integration von Netzen über ähnlichem Grundwortschatz in verschiedenen Sprachen

- ◆ Beispiel: EuroWordNet – *Interlingual Index*

- ◆ Wenig Automatisierungsmöglichkeit!
- ◆ Indirektes Verknüpfen auf Referenzgröße

- ◆ Wichtige Eigenschaften

- ◆ stabile Referenzgröße
- ◆ dokumentierte Grob-Ontologie
- ◆ Nicht jedes Element des Netzes muss verknüpft sein



Indirektes Verknüpfen

Szenario III: Teil-Netze

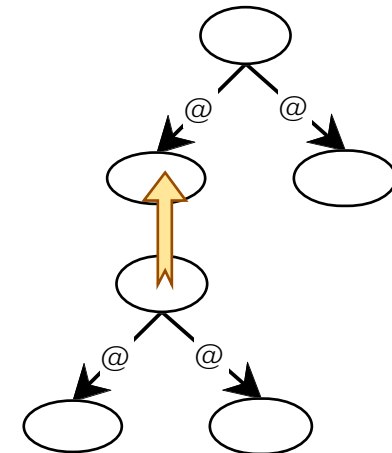
Integration von spezifischeren Netzen in allgemeinere

◆ Beispiele

- ◆ Manuelles Einhängen kompakter Netzteile
- ◆ Halbautomatisches Integrieren von Netzen [Magnini et al. 02]
- ◆ Automatische Integration von zusätzlicher Terminologie [Buitelaar et al. 02]

◆ Wichtige Eigenschaften

- ◆ Ontologische Tiefe des spezifischeren Netzes
- ◆ Gemeinsame Lexik



Identifizieren von Anknüpfungspunkten

Fazit: GermaNet und UniNet



Kritische Kriterien

- ? Gemeinsame Lexik
- ? Strukturähnlichkeit
- ? Stabilität und ontologische Explizitheit von GermaNet

Idee

- ◆ Partielles indirektes Indizieren des UniNets auf ILIs, welche mit GermaNet verknüpft sind.
- ◆ *inter-lingual index* wird als *inter-net index* verwendet.

Probleme

- ◆ Verfügbarkeit der GermaNet-seitigen ILI
 - ◆ ...
- 

Schluss



Oder

- ◆ Lösen sich die klassischen semantischen Netze sowieso in moderneren Formen und Formaten der semantischen Repräsentation ontologischer, begrifflicher und terminologischer Begebenheiten auf?