Automatic Construction of Linguistic Resources

Sebastian Blohm

Outline of Talk

> **Introduction:** Linguistic Resources in Human - Computer Interaction
> **Experiment 1:** Word Categories (Riloff, Shepherd)
> **Experiment 2:** Refining Taxonomies (Ide, Veronis)
> **Experiment 3:** Extraction and Structuring of Information (Embley, Capbell, Smith)
> **Document search** with the help of semantic information
> **Conclusion/Discussion**

totally: 29 slides

Typen von semantischem Wissen

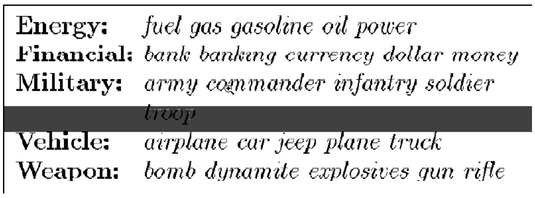
> **Wissen über die Zuordnung zwischen Repräsentation und Fakten**
> **Wissen um mögliche** Inferenzen auf der **Repräsentation.**

Experiment 1

A Corpus-Base Approach for Building Semantic Lexicons.

Riloff, E. and Shepherd, J (1997). Saltlake City.

Seedwords

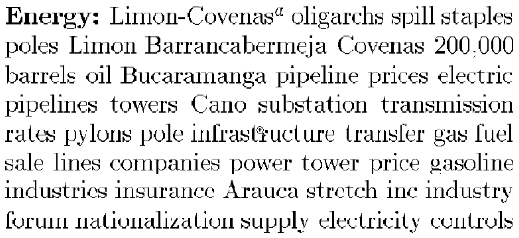


Energy:	*fuel gas gasoline oil power*
Financial:	*bank banking currency dollar money*
Military:	*army commander infantry soldier troop*
Vehicle:	*airplane car jeep plane truck*
Weapon:	*bomb dynamite explosives gun rifle*

Bootstrapping Algorithmus

```
wähle eine Kategorie C
init seedwords
categorywords={}
wiederhole ca. 8 Mal
   finde alle Kontexte der Seedwords
   für alle Wörter W, die in den Kontexten
     auftreten:
   Score(W,C) = freq(W im Kontext von C) / freq(W)
   categorywords = categorywords + (W,Score(W,C))
        kopiere die 5 besten W in seedwords
```

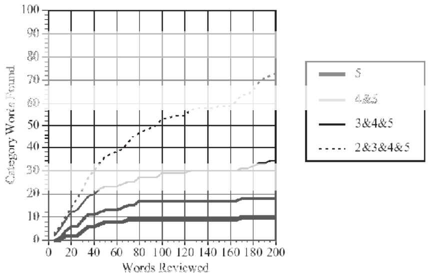
Words in the Category



Energy: Limon-Covenas[a] oligarchs spill staples poles Limon Barrancabermeja Covenas 200,000 barrels oil Bucaramanga pipeline prices electric pipelines towers Cano substation transmission rates pylons pole infrastructure transfer gas fuel sale lines companies power tower price gasoline industries insurance Arauca stretch inc industry forum nationalization supply electricity controls

Measuring Success

5: core member of category
1: not related



Experiment 2

Refining Taxonomies Extracted from Machine Readable Dictionaries. Ide, N. and Véronis, J. (1994). New York

Typen semantischer Beziehungen

> zwischen Wortformen (lexical)

Synonym: zwei Worte haben die gleiche Bedeutung.

> zwischen Wortbedeutungen (semantic)

Hyperonym, Hyponym: Oberbegriff/Spezialisierung
Antonym: Gegenteil
Meronym/Holonym: Bestandteil/ entsteht aus

Mögliche Eigenschaften dieser Beziehungen:

» transitiv
» symmetrisch oder reziprok

Die Beziehungen eigenen sich zum Aufbau einer **Taxonomie** in Graphnotation.

Beispielnyme

Xnyme des Wortes **Fußballspieler**:

Hyperonym -	Sportler
Hyponym -	Torwart
Meronym –	Trikot
Holonym -	Mannschaft
Antonym -	*Angler
Synonym -	...

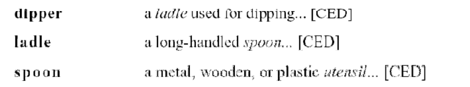
Symmetrie : antonym(heiß,kalt)<->antonym(kalt,heiß)

Reziprozität: hyper(Sportler,Fussballspieler)<-> hypo(Fussballspieler,Sportler)

Transitivität: hyper(Fussballspieler,Torwart) && hyper(Sportler,Fussballspieler) -> hyper(Sportler,Torwart)

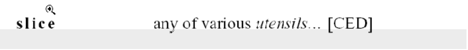
Extracting Semantic Relations

> The hypernym is head of the defining noun phrase



dipper	a *ladle* used for dipping... [CED]
ladle	a long-handled *spoon*... [CED]
spoon	a metal, wooden, or plastic *utensil*... [CED]

> or alternatively head of a phrase starting with 'of'.



slice	any of various *utensils*... [CED]

Problems with Taxonomie Extraction

> **Attachment too high** – description uses a hypernym of the actual hypernym.

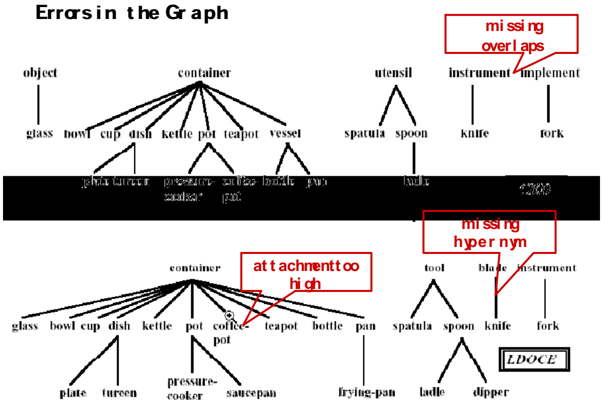
> **Missing hypernyms** – other way of description used.

> **Missing overlaps:** - out of several possible hypernyms only one is given.

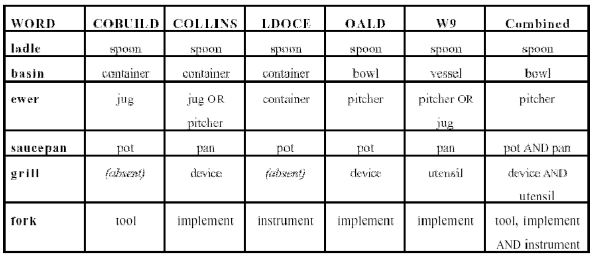
> **Or-conjoined heads:** - description yields several heads not all being a universal hypernym

> **Circularity** – words are described by each other.

Errors in the Graph

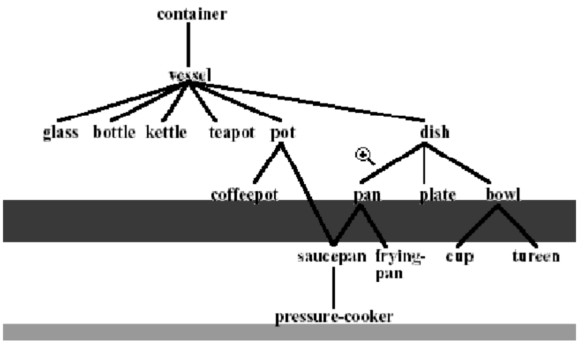


Merging hierarchies



WORD	COBUILD	COLLINS	LDOCE	OALD	W9	Combined
ladle	spoon	spoon	spoon	spoon	spoon	spoon
basin	container	container	container	bowl	vessel	bowl
ewer	jug	jug OR pitcher	container	pitcher	pitcher OR jug	pitcher
saucepan	pot	pan	pot	pot	pan	pot AND pan
grill	*(absent)*	device	*(absent)*	device	utensil	device AND utensil
fork	tool	implement	instrument	implement	implement	tool, implement AND instrument

Corrected Graph



Experiment 3

Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents.

Ebley, D., Campbell, D., Smith, R. (1998)

Raw data

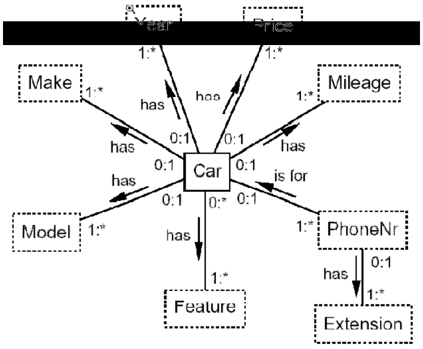
Criteria:

> „data-richness“

> „narrow ontological breadth“

```
'96 CHEV Monte Carlo Z34, loaded, bright Red
15,000 actual miles!  A great buy at $14,990,
$750 to 1000 down.  MURDOCK CHEVROLET 298-8090
#####
'94 CHEV Corsica, 88,281 miles.  Ask for #16. $4,900.
Government Surplus533-5885
#####
'89 AUDI 80, red, auto., p/w, p/l, sunroof, loaded, 126K,
new trans., new diff.  Runs perfect, must sell, $3300 obo.
gcall Nate, 554-4414
```

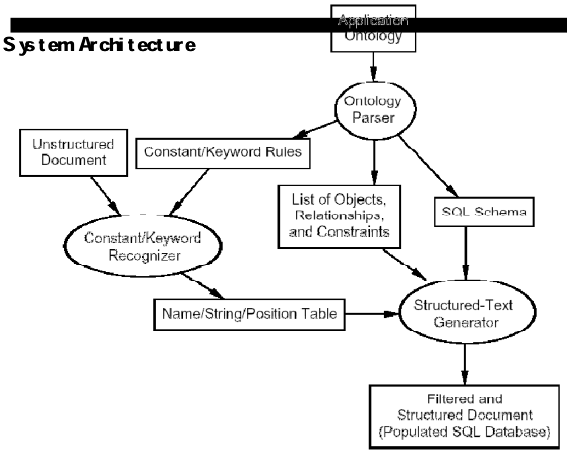
Graphical Ontology



Ontology File

```
Car [0:1] has Year [1:*];
Year {regexp[2]: "\d{2} : ([^\$\d]|^)\d{2}[^,\dkK]",
        "\d{2} : ([^\$\d]|^)\d{2},[^\d]",
        "\d{2} : \b'\d{2}\b" };
Car [0:1] has Make [1:*];
Make {regexp[10]: "\bchev\b", "\bchevy\b", ... };
Car [0:1] has Model [1:*];
Model {regexp[16]: "88 : \bolds\S*\s*88\b",
        "80 : \baudi\S*\s*80\b", "\bacclaim\b", ... };
Car [0:1] has Mileage [1:*];
Mileage {regexp[8]: "\b[1-9]\d{1,2}k",
        "[1-9]\d?,\d{3} : [^\$\d][1-9]\d?,\d{3}[^\d]" }
        {context: "\bmiles\b", "\bmi\.", "\bmi\b"};
Car [0:*] has Feature [1:*];
Feature {regexp[20]:
        -- Colors
                "\baqua\s+metallic\b", "\bbeige\b", ...
        -- Transmission
                "(5|6)\s*spd\b", "auto : \bauto(\.|,)",
        -- Accessories
                "\broof\s+rack\b", "\bspoiler\b", ...
```

System Architecture


[redacted]
Ontology Parser
Unstructured Document
Constant/Keyword Rules
List of Objects, Relationships, and Constraints
SQL Schema
Constant/Keyword Recognizer
Name/String/Position Table
Structured-Text Generator
Filtered and Structured Document (Populated SQL Database)

name/string/position table

Year|96|2|3
Make|CHEV|5|8
Model|Monte Carlo|10|20
~~Year|34|23|24~~
Feature|Red|42|44
Mileage|15,000|46|51
KEYWORD(Mileage)|miles|60|64
Price|14,990|84|89
~~Price|750|93|95~~
~~Mileage|1000|100|103~~
~~Make|CHEVROLET|120|128~~
PhoneNr|298-8090|130|137

extracted information

```
Year Make       Model            Price
---- ---------- ---------------- --------
94   DODGE                       4,995
94   DODGE      Intrepid         10,000
[redacted]
90   FORD       Probe
88   FORD       Escort           1000
```

Example of application

> Using Wordnet in Text retrieval

Die Wordnet – Datenbank: Browser

Zu Nomen kann der Browser Anzeigen:

> Synonyme nach Häufigkeit

> Synonyme nach Ähnlichkeit

> „Coordinate Terms“ – Alternative Begriffe des gleichen Hyperonyms

> Hyperonyme und Hyponyme

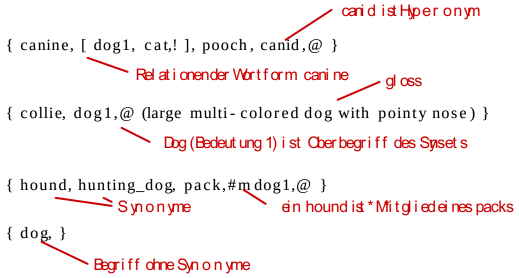
> Bekanntheit (abgeschätzt durch Polysemie)

Zu den Relationen ist auch eine für Menschen lesbare Definition (**gloss**) gespeichert.

WordNet Browser.lnk

Die Wordnet – Datenbank: Datenstruktur

Dies ist die Datenstruktur von Lexicographer-Files, die manuell erstellt werden.



canid ist Hyperonym

{ canine, [dog1, cat,!], pooch, canid,@ }

Relationen der Wortform canine

gloss

{ collie, dog1,@ (large multi-colored dog with pointy nose) }

Dog (Bedeutung 1) ist Oberbegriff des Synsets

{ hound, hunting_dog, pack,#m dog1,@ }

Synonyme

ein hound ist * Mitglied eines packs

{ dog, }

Begriff ohne Synonyme

Using WordNet in Textretrieval

> Text-Suche als Vektorraummodell

Mehrdeutigkeit bei der Suche:

> **Homographen** verringern Precision

> **Synonyme** verringern Recall

Concept Matching:

> Queries und Dokumente als semantic concordances

> Semantic Tagging mit Hilfe von „Hoods“

> automatisches „Semantic Tagging“ insb. bei Query ein Problem

Query Expansion:

> Synsets aus der Umgebung werden der hinzugefügt.

→beide Verfahren scheitern am „Semantic Tagging“

Conclusion and discussion

References

Riloff, E. and Shepherd, J (1997). A Corpus-Base Approach for Building Semantic Lexicons. Saltlake City.

Refining Taxonomies Extracted from Machine Readable Dictionaries. Ide, N. and Véronis, J. (1994). New York.

Ebley, D., Campbell, D., Smith, R. (1998). Ontology -Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. Provo.

http://www.cogsci.princeton.edu/~wn/

» Insbesondere „Five Papers on Wordnet“

Fellbaum, C., (1998). WordNet, an electronical lexical database . Cambride, MA: MIT Press.

» Insbesondere: Voorhees E. M.. Using WordNet for Text Retrieval.