Martha Palmer
University of Pennsylvania

Paul Kingsbury
University of Pennsylvania

Daniel Gildea
University of Rochester

The Proposition Bank: An Annotated Corpus of Semantic Roles

Rositsa Nedyalkova

Propbank in Verbindung zu den vorigen Sitzungen

Thematische Rollen→semantische Rollen

Framenet → Vergleich

Verbalternationen → in Propbank vertreten

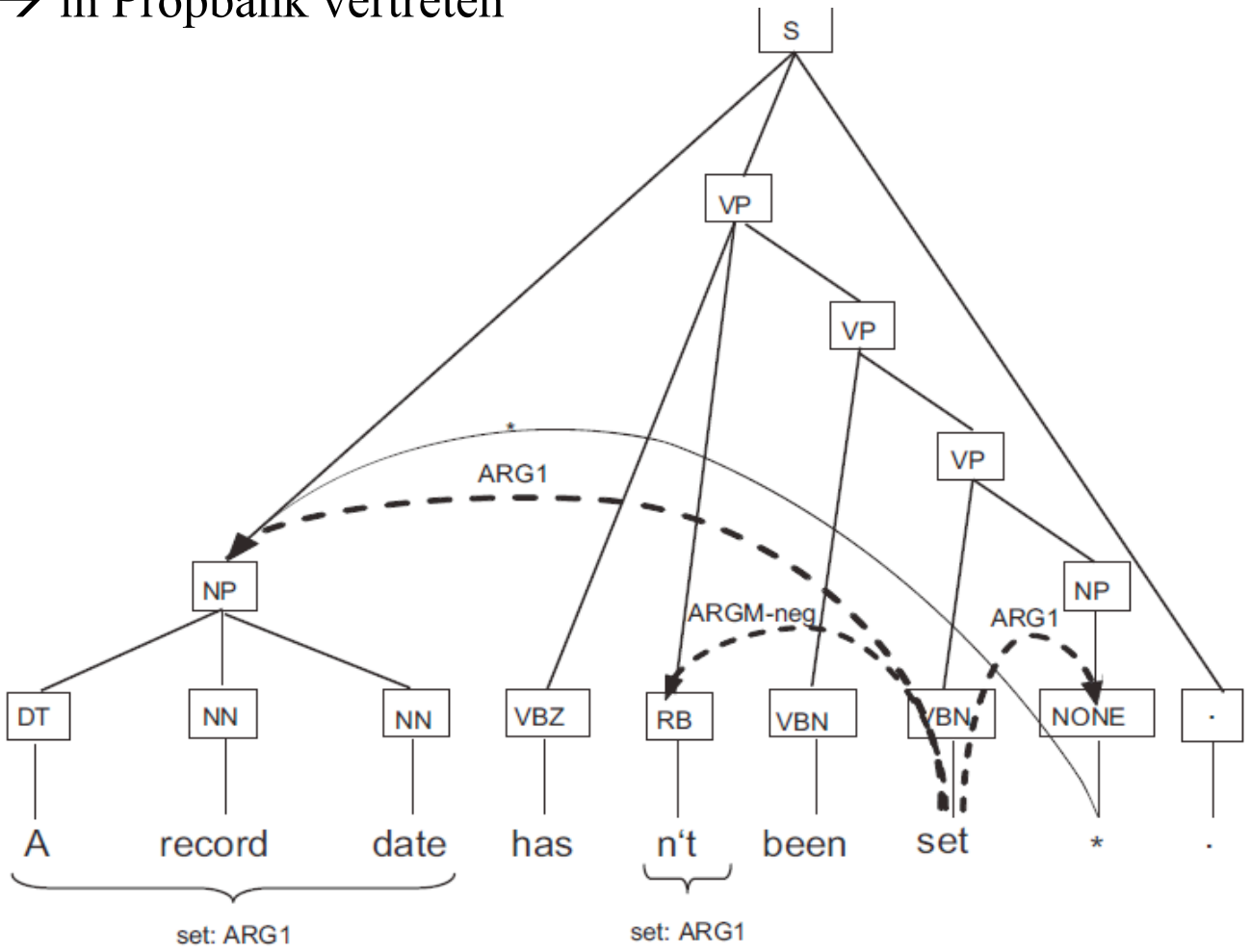


Figure 2: The phrase structure representation of sentence wsj02wsj_0202 in PropBank

Einführung

- Was ist Propbank?

- PennTreebank
- Levin Klassen

- Ziel:durch Verbalternationen ein nützliches Niveau im Bereich der semantischen Repräsentation zu erreichen.
- Potentielle Anwendung: Extrahierung von Information ,Fragen beantworten,Maschinenübersetzung.

Inhalt

1.Semantische Rollen und Syntaktische Alternation

2.Annotation Schemata:Aussuchen des Sets semantischer Rollen

3.Propbank - Entwicklungsprozess

4.Propbank und Framenet

5.Quantitative Analyse der semantischen Rollen Labels

6.Automatische Ermittlung der semantischen Rollen Labels

7.Fazit

Einführung

(1) John broke the window.

(2) The window broke.

- . . . [$_{Arg0}$ the company] to . . . *offer* [$_{Arg1}$ a 15% to 20% stake] [$_{Arg2}$ to the public] (wsj_0345)[1]

  . . . [$_{Arg0}$ Sotheby's] . . . *offered* [$_{Arg2}$ the Dorrance heirs] [$_{Arg1}$ a money-back guarantee] (wsj_1928)

  . . . [$_{Arg1}$ an amendment] *offered* [$_{Arg0}$ by Rep. Peter DeFazio] . . . (wsj_0107)

  . . . [$_{Arg2}$ Subcontractors] will be *offered* [$_{Arg1}$ a settlement] . . . (wsj_0187)

- Die Propbank konzentriert sich auf die Argumentenstruktur der Verben und entwickelt ein komplett mit semantischen Rollen annotiertes Korpus ,einschließlich Rollen,betrachtet als Argumente und Attribute.
- Die Propbank erlaubt uns die Häufigkeit der syntaktischen Variationen zu bestimmen und dadurch auch die Probleme zu lösen, die sie bei den natürlichen Sprachen bereiten.

1.Semantische Rollen und syntaktische Alternation

- Verbnet und Propbank:

Verbnet Klassen : Die Basis sind die Levin Klassen

Hinzufügen von Subklassen,die syntaktischen und semantischen Zusammenhang zwischen den Elementen einer Klassen bilden.

Jede Verbnet Klasse wird beschrieben durch:

*thematische Rolle

*Argumentenabgrenzung

*Frames:syntaktische Beschreibung + semantische Prädikate(motion,contact,cause)

Class Hit-18.1			
Roles and Restrictions: Agent[+int_control] Patient[+concrete] Instrument[+concrete]			
Members: bang, bash, hit, kick, ...			
Frames:			
Name	Example	Syntax	Semantics
Basic Transitive	Paula hit the ball	Agent V Patient	cause(Agent, E)manner(during(E), directedmotion, Agent) ! contact(during(E), Agent, Patient) manner(end(E),forceful, Agent) contact(end(E), Agent, Patient)

2.Annotatonsschemata:Bestimmung des Sets semantischer Rollen

- Semantische Rollen auf der Verb-by-Verb Basis

Die individuellen semantischen Argumente eines Verbs sind nummeriert,beginnend mit der Null:

Arg0= prototypischer Agent

Arg1 = prototypischer Patient(Dulder)

- Zusätzliche auf das Verb bezogene Rollen

Beispiel: die Verben ***accept*** und ***kick***

Frameset accept.01 ‘‘take willingly’’
Arg0: Acceptor
Arg1: Thing accepted
Arg2: Accepted-from
Arg3: Attribute
Ex:[Arg0 He] [ArgM-MOD would][ArgM-NEG n’t] accept [Arg1 anything of value] [Arg2 from those he was writing about]. (wsj_0186)

Frameset kick.01 ‘‘drive or impel with the foot’’
Arg0: Kicker
Arg1: Thing kicked
Arg2: Instrument (defaults to foot)
Ex2: [Arg0 John] tried to kick [Arg1 the football], but Mary pulled it away at the last moment.

2.Annotatonsschemata:Bestimmung des Sets semantischer Rollen

- Roleset
- Frameset = Roleset + assozierte Frames
- Frames file = Die Sammlung aller Frameset Einträge eines Verbs.

- Attributsets(ArgMs)

Table 1
Subtypes of the ArgM modifier tag.

LOC: location	CAU: cause
EXT: extent	TMP: time
DIS: discourse connectives	PNC: purpose
ADV: general purpose	MNR: manner
NEG: negation marker	DIR: direction
MOD: modal verb	

2.Annotatonsschemata:Bestimmung des Sets semantischer Rollen

2.1 Charakterisierung von Framesets-basiert auf Semantik und Syntax

- Wenn ein Verb zwei verschiede Bedeutungen hat,die nicht gleiche Anzahl von Argumenten benötigen,hat das Verb zwei Framesets.

Frameset **decline.01** "go down incrementally"

Arg1: entity going down

Arg2: amount gone down by, EXT

Arg3: start point

Arg4: end point

Ex: . . . [$_{\text{Arg1}}$ its net income] *declining* [$_{\text{Arg2-EXT}}$ 42%] [$_{\text{Arg4}}$ to $121 million] [$_{\text{ArgM-TMP}}$ in the first 9 months of 1989]. (wsj_0067)

Frameset **decline.02** "demure, reject"

Arg0: agent

Arg1: rejected thing

Ex: [$_{\text{Arg0}}$ A spokesman$_i$] *declined* [$_{\text{Arg1}}$ *trace*$_i$ to elaborate] (wsj_0038)

2.Annotatonsschemata:Bestimmung des Sets semantischer Rollen

- Verben,deren Alternationen(kausativ-inchoativ,Ausfallen des Objekts,transitiv-intransitiv)die Bedeutung nicht ändern,haben nur ein Frameset und die Alternationen beziehen sich auf das eine Frameset,wo manche Argumente nicht genau angegeben sind.

Frameset **open.01** "cause to open"

Arg0: agent

Arg1: thing opened

Arg2: instrument

Ex1: [$_{Arg0}$ John] *opened* [$_{Arg1}$ the door]

Ex2: [$_{Arg1}$ The door] *opened*

Ex3: [$_{Arg0}$ John] *opened* [$_{Arg1}$ the door] [$_{Arg2}$ with his foot]

2.Annotatonsschemata:Bestimmung des Sets semantischer Rollen

2.2 sekundäre Behauptungen

- EXT(extent)
- PRD(secondary prediction)

Mary called John an idiot. (PRD) Mary called John a taxicab. (no PRD)

Das Verb “call” in der Bedeutung von “attach a label to”

predicative reading	*ditransitive reading*
Mary called John a doctor.	Mary called John a doctor.[5]
(LABEL)	(SUMMON)
Arg0: Mary	Arg0: Mary
Rel: called	Rel: called
Arg1: John (item being labeled)	Arg2: John (benefactive)
Arg2-PRD: a doctor (attribute)	Arg1: a doctor (thing summoned)

2.Annotatonsschemata:Bestimmung des Sets semantischer Rollen

2.3Zusammengefasste Argumente

Frameset **hit** "strike"

Arg0: hitter

Arg1: thing hit, target

Arg2: instrument of hitting

Ex3: All arguments: [$_{Arg0}$ John] *hit* [$_{Arg1}$ the tree] [$_{Arg2}$ with a stick].

kick und **hammer**

2.4 Role Labels und syntaktische Bäume

Die Knoten der syntaktischen Bäume der Penn Treebank werden durch semantischen Rollen annotiert.

Man darf die Struktur der Bäume nicht ändern

→ Bei manchen Ausdrücken entstehen Annotationsprobleme wie bei:

Präpositionalphrasen:

John poured the water into the bottle.

1.Wenn ein Argument als Zie definiert ist dann**:**

Zielort von Wasser = Flasche(nicht “in die Flasche”) Präpositionsphrase – in die Flasche

Also soll dieses Argument mit der NP assoziert werden

2.ArgMs,die Präpositionsphrase darstellen,sollen auf dem PP Niveau annotiert werden.

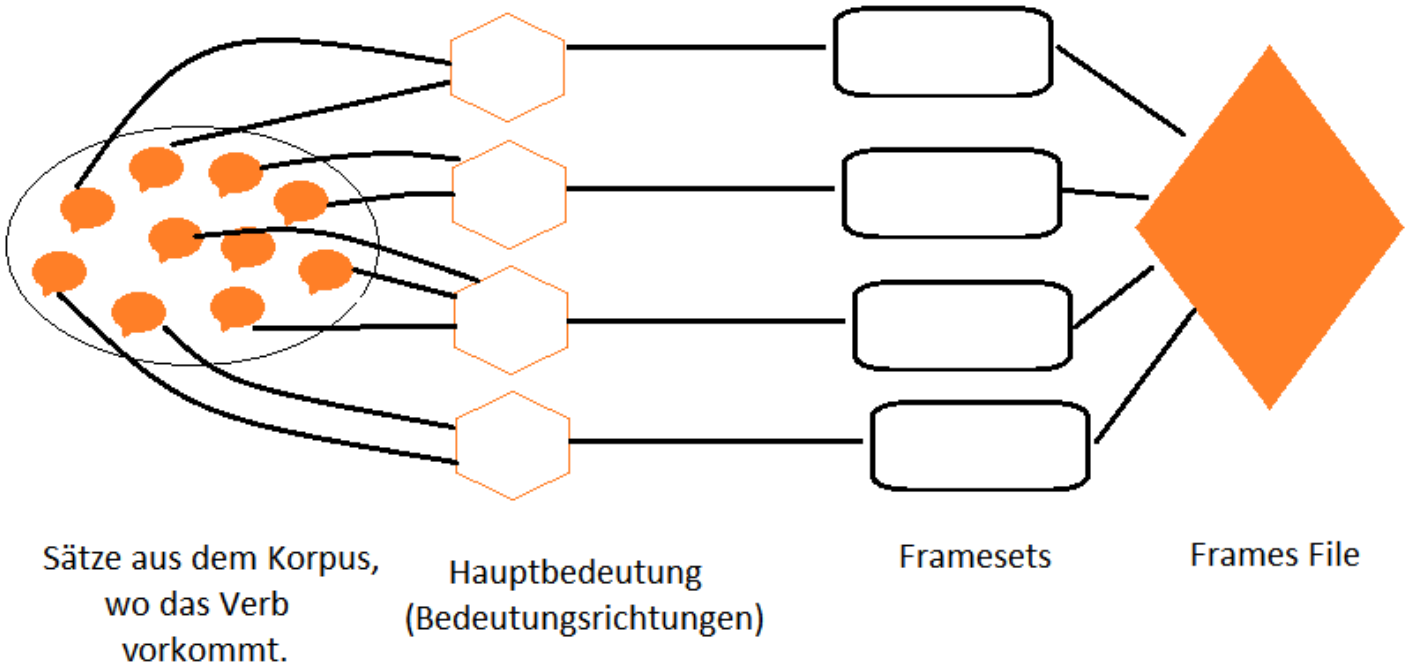
→Um Konflikte zu vermeiden,werden nummerierte Argumente auf beiden Levels annotiert.

Weitere Beispiele dafür:

Out of,according to

3.Propbank - Entwicklungsprozess

3.1 Framing



3.Propbank - Entwicklungsprozess

3.2 Annotation

I. Regelbasierter Argument Tagger
II. Handannotation
III. Kappa Statistik(Siegel und Castellan 1988)

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) = Wahrscheinlichkeit der Kongruenz(zwischen den verschiedenen Annotationen)

P(E) = zufälliges Agreement

- Rollenidentifikation (role vs nonrole)
- Rollenklasifikation (Arg0 vs Arg 1 usw)

Interannotator agreement.

		$P(A)$	$P(E)$	κ
Including ArgM	Role identification	.99	.89	.93
	Role classification	.95	.27	.93
	Combined decision	.99	.88	.91
Excluding ArgM	Role identification	.99	.91	.94
	Role classification	.98	.41	.96
	Combined decision	.99	.91	.93

4.Framenet und PropBank

+ Gemeinsames Ziel: Die syntaktische und semantische Realisation der Argumente eines Prädikats zu dokumentieren.

+ Gemeinsames Mittel:Durch semantischen Rollen annotieres Korpus.

- Unterschiedliche Methodik:

❖ Framenet

- Frame-by-Frame Basis :semantische Frames(z.B Commerce);Frame Elemente(Buyer,Goods,Seller,Money);Frame(buy,sell usw..)

-Korpus:British National Corpus;Sätze,wo die Realisation derArgumente enfach dargestellt wird, werden bevorzugt.

❖ PropBank

- Korpus: Wall Street Journal,geerbt von Penn Treebank

(Propbank soll alle Absätze annotieren,die in Penn Treebank vorkommen,egal wie komplex sie sind)

4.Framenet und PropBank

Comparison of frames.

PropBank		FrameNet
buy	*sell*	COMMERCE
Arg0: buyer	Arg0: seller	Buyer
Arg1: thing bought	Arg1: thing sold	Seller
Arg2: seller	Arg2: buyer	Payment
Arg3: price paid	Arg3: price paid	Goods
Arg4: benefactive	Arg4: benefactive	Rate/Unit

PropBank annotation:

[$_{\text{Arg0}}$ Chuck] *bought* [$_{\text{Arg1}}$ a car] [$_{\text{Arg2}}$ from Jerry] [$_{\text{Arg3}}$ for $1000].

[$_{\text{Arg0}}$ Jerry] *sold* [$_{\text{Arg1}}$ a car] [$_{\text{Arg2}}$ to Chuck] [$_{\text{Arg3}}$ for $1000].

[$_{\text{Arg1}}$ A car] was *bought* [$_{\text{Arg0}}$ by Chuck].

[$_{\text{Arg1}}$ A car] was *sold* [$_{\text{Arg2}}$ to Chuck] [$_{\text{Arg0}}$ by Jerry].

[$_{\text{Arg2}}$ Chuck] was *sold* [$_{\text{Arg1}}$ a car] [$_{\text{Arg0}}$ by Jerry].

FrameNet annotation:

[$_{\text{Buyer}}$ Chuck] *bought* [$_{\text{Goods}}$ a car] [$_{\text{Seller}}$ from Jerry] [$_{\text{Payment}}$ for $1000].

[$_{\text{Seller}}$ Jerry] *sold* [$_{\text{Goods}}$ a car] [$_{\text{Buyer}}$ to Chuck] [$_{\text{Payment}}$ for $1000].

[$_{\text{Goods}}$ A car] was *bought* [$_{\text{Buyer}}$ by Chuck].

[$_{\text{Goods}}$ A car] was *sold* [$_{\text{Buyer}}$ to Chuck] [$_{\text{Seller}}$ by Jerry].

[$_{\text{Buyer}}$ Chuck] was *sold* [$_{\text{Goods}}$ a car] [$_{\text{Seller}}$ by Jerry].

5.Quantitative Analyse - Semantic Role Labels

Darstellung eines Systems,das die semantischen Rollen automatisch bestimmt.

(Merlo and Stevenson, 2001)

5.1. Charakterisierung der semantischen Rollen durch ihre syntaktische Realisierung

- Wie häufig kommt eine bestimmte semantische Rolle in spezifischen syntaktischen Position vor?

  -Zur Vereinfachung werden Partizip Präsens und Perfekt nicht berücksichtigt.

  -Regeln:

*jede NP unter S in der Treebank = syntaktisches Subjekt.(Arg0)

*jede NP unter VP = syntaktisches Objekt.(Arg1)

In allen anderen Fällen werden die syntaktischen Kategorien übernommen,die an den Knoten der Treebank Bäume bestimmt sind.

5.Quantitative Analyse - Semantic Role Labels

Most frequent semantic roles for each syntactic position.

Position	Total	Four most common roles (%)								Other roles (%)
Sub	37,364	Arg0	79.0	Arg1	16.8	Arg2	2.4	TMP	1.2	0.6
Obj	21,610	Arg1	84.0	Arg2	9.8	TMP	4.6	Arg3	0.8	0.8
S	10,110	Arg1	76.0	ADV	8.5	Arg2	7.5	PRP	2.4	5.5
NP	7,755	Arg2	34.3	Arg1	23.6	Arg4	18.9	Arg3	12.9	10.4
ADVP	5,920	TMP	30.3	MNR	22.2	DIS	19.8	ADV	10.3	17.4
MD	4,167	MOD	97.4	ArgM	2.3	Arg1	0.2	MNR	0.0	0.0
PP-in	3,134	LOC	46.6	TMP	35.3	MNR	4.6	DIS	3.4	10.1
SBAR	2,671	ADV	36.0	TMP	30.4	Arg1	16.8	PRP	7.6	9.2
RB	1,320	NEG	91.4	ArgM	3.3	DIS	1.6	DIR	1.4	2.3
PP-at	824	EXT	34.7	LOC	27.4	TMP	23.2	MNR	6.1	8.6

5.Quantitative Analyse - Semantic Role Labels

- 5.2.Verbindung der Verb Klassen mit den spezifischen syntaktischen Konstruktionen

Unergative: [$_{\text{Causal Agent}}$ The jockey] *raced* [$_{\text{Agent}}$ the horse] past the barn.

[$_{\text{Agent}}$ The horse] *raced* past the barn.

Unaccusative: [$_{\text{Causal Agent}}$ The cook] *melted* [$_{\text{Theme}}$ the butter] in the pan.

[$_{\text{Theme}}$ The butter] *melted* in the pan.

Object-Drop: [$_{\text{Agent}}$ The boy] *played* [$_{\text{Theme}}$ soccer].

[$_{\text{Agent}}$ The boy] *played*.

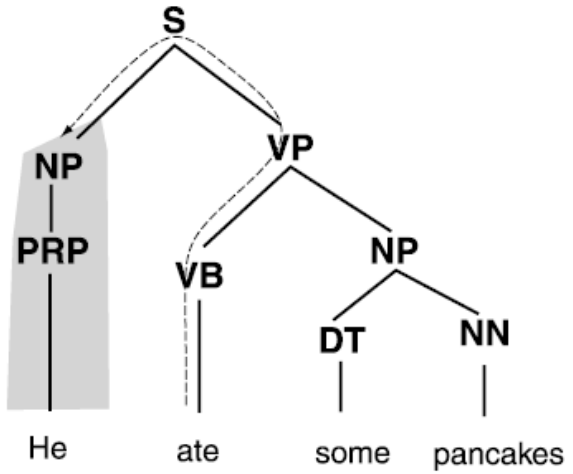
6.Automatische Ermittlung der Semantischen Rollen

6.1Das statistische System von Gildea und Jurafsky(2002) - trainert auf die Daten von Framenet zur automatischen Bestimmung der semantische Rollen.

6.1.1.Das System

Die Wahrscheinlichkeiten einer Parse Konstituente,die zu einer semantischen Rolle gehört,werden durch die folgenden Merkmale beschrieben.

- Typ der Phrase: (NP,VP,S)
- Parse Baumpfad
- Position
- Voice
- Lemma(headword)



Prädikat→ Argument : VB↑VP↑S↓

ate He

6.Automatische Ermittlung der Semantischen Rollen

- Das System bestimmt die Argumentenrollen im Datenbank durch:

$\text{argmax}_{r_{1\ldots n}} P(r_{1\ldots n}|F_{1\ldots n}, p)$

argmax: die höchste WK-Belegung der Rolen r_i zu allen Konstituenten i im Satz.

$F_i = \{pt_i, path_i, pos_i, v_i, h_i\}$ Merkmalset für jedes i und Prädikat p

I. Wir teilen die WK-Bewertung in 2 Teilen: $P(r_i|F_i)$ und $P(r_i|p)$ $P(r_i \mid F_i, p)$

II. Diese WK-en werden kombiniert mit: $P(\{r_{1\ldots n}\}|p)$

→ $P(r_{1\ldots n}|F_{1\ldots n}, p) \approx P(\{r_{1\ldots n}\}|p) \prod_i \frac{P(r_i|F_i, p)}{P(r_i|p)}$

Set von den Rollen im Satz|Prädikat

6.Automatische Ermittlung der Semantischen Rollen

- Resultate

  Das selbe System wurde auf eine vorläufige Version von Propbank angewendet.

  - Aus dem Propbank Datenset: 72,109 Prädikat-Argument Strukturen(190,815 individuelle Argumente und Beispiele von 2,462 lexikalischen Prädikaten + die Annotationen einer Sektion von der Treebank.

Das System wurde unter zwei Bedingungen getestet.

1. Es sind Konsituenten vorgegeben,die Argumente zu den Prädikaten sind,und das System soll nur die korrekte Role bestimmen.
2. Das System soll die Argumente in den Sätzen finden und sie richtig annotieren.

Die Resultate sind sehr ähnlich wie diese bei Framenet

6.Automatische Ermittlung der Semantischen Rollen

Accuracy of semantic-role prediction (in percentages) for known boundaries (the system is given the constituents to classify).

	Accuracy		
	FrameNet	PropBank	PropBank > 10 examples
Automatic parses	82.0	79.9	80.9
Gold-standard parses		82.0	82.8

Accuracy of semantic-role prediction (in percentages) for unknown boundaries (the system must identify the correct constituents as arguments and give them the correct roles).

	FrameNet		PropBank		PropBank > 10 examples	
	Precision	Recall	Precision	Recall	Precision	Recall
Automatic parses	64.6	61.2	68.6	57.8	69.9	61.1
Gold-standard parses			74.3	66.4	76.0	69.9
Gold-standard with traces			80.6	71.6	82.0	74.7

7.Fazit

Syntaktische Strukturen + Semantische Rolen = Detailierte Semantische Repräsentation

Penn Treebank Levin Propbank

- Die Annotation der Propbank – bei dem Training von automatischen Taggers angewendet

+ Entwicklung der IE Systeme(Information Extraction)

+Maschinenübersetzung

Weitere Projekte:

Propbank für die Chinesische (Xue und Palmer 2003) und Koreanische Sprache

New York University -> Annotation der Nominalisierung

Erweiterung der Propbank

Kipper,Palmer,und Rambow) – Neue Version mit mehr informativen thematischen Labels,basierend auf VerbNet

Gemeinsame Arbeit mit Framenet – Mapping zwischen den beider Annotationen