

Semantic Coordination mit CTXMATCH

Jorgen Schäfer

Seminar Semantic Integration

Arbeitsbereich Wissens- und Sprachverarbeitung
Universität Hamburg

Veranstalter: Özgür Özçep, Carola Eschenbach
Sommersemester 2006

7. Juli 2006

1 Einführung und Motivation

2 CTXMATCH

- Überblick
- Arbeitsweise

3 Gütekriterien

4 Zusammenfassung

1 Einführung und Motivation

2 CTXMATCH

- Überblick
- Arbeitsweise

3 Gütekriterien

4 Zusammenfassung

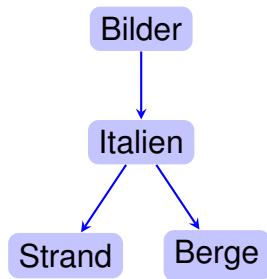
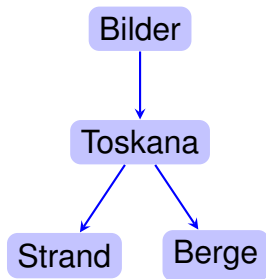
Semantic Coordination

- Semantic Coordination findet semantische Beziehungen zwischen Knoten abstrakter Strukturen
- Weitere Namen
 - *Schema Matching*
 - *Ontology Mapping*
 - *Semantic Integration*
 - ...

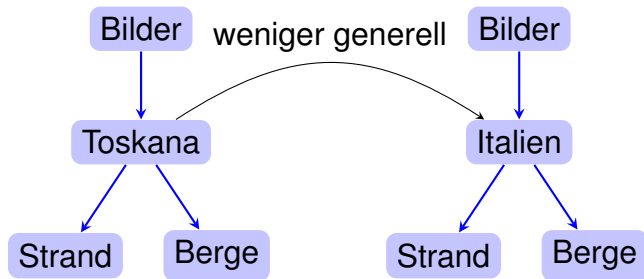
Hierarchical Classifications (HC)

- Insbesondere hier: Hierarchical Classifications (HC)
- HCs sind meist Strukturen, deren Beschriftung aus der Sprache der Benutzer stammt
- Die Semantik dieser Beschriftungen ist relevant
- Beispiele
 - Google™ Directory, Yahoo!™ Directory
 - Dienstleistungsregister (UDDI)
 - Märkte (Waren werden klassifiziert)
 - Hierarchische Dateisysteme
 - ...
- Ziel: Paarweise Relationen zwischen allen Knoten zweier HCs

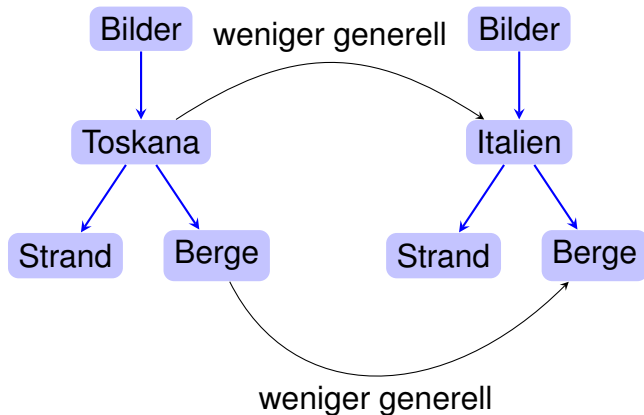
Bilder von Bergen



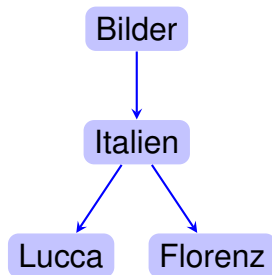
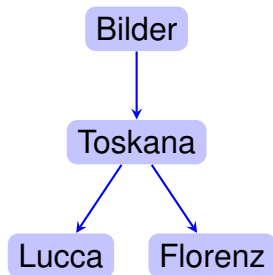
Bilder von Bergen



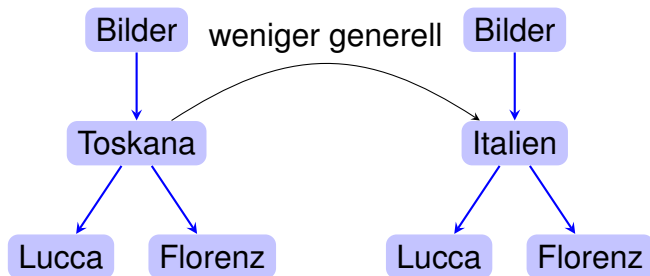
Bilder von Bergen



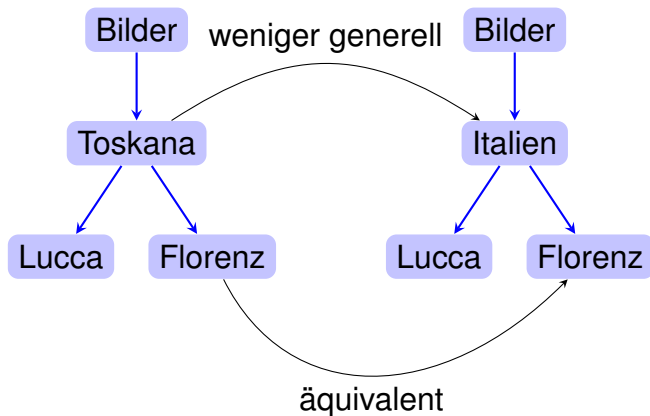
Bilder von Städten



Bilder von Städten



Bilder von Städten



- Menschen verstehen hier *intuitiv*
 - Berge₁ ist *weniger generell* als Berge₂,
denn *wir wissen*: Berge-in-der-Toskana \subset Berge-in-Italien
 - Aber: Florenz₁ ist *äquivalent* zu Florenz₂,
denn *wir wissen*: Florenz-in-der-Toskana \equiv Florenz-in-Italien
- Also eine unterschiedliche Relation trotz gleicher Struktur
- D.h. die Struktur alleine ist nicht aussagekräftig
- Daher benötigen wir die Semantik der Begriffe

Semantik von Beschriftungen

- Beschriftungen aus einer bekannten Sprache
- Erlaubt Benutzern die implizite semantische Koordinierung mithilfe dieser Sprache
- Diese semantische Information kann durch Wörterbücher verfügbar gemacht werden

1 Einführung und Motivation

2 CTXMATCH

- Überblick
- Arbeitsweise

3 Gütekriterien

4 Zusammenfassung

1 Einführung und Motivation

2 CTXMATCH

- Überblick
- Arbeitsweise

3 Gütekriterien

4 Zusammenfassung

CtxMatch: Idee

- Genereller Algorithmus zur Koordinierung von Strukturen mit natürlichsprachlichen Beschriftungen
- Für spezielle Typen von Strukturen müssen einige Teile angepasst werden
- Hier: HCs – Hierarchical Classifications
- Semantische Informationen aus WORDNET

CtxMatch: Funktionsweise

- Eingabe: Zwei HCs, H und H'
- Ausgabe: Die semantische Relation für jedes Knotenpaar $(k, k') : k \in H$ und $k' \in H'$
- Mögliche Relationen in HCs:
 - k äquivalent-zu k'
 - k genereller-als k'
 - k weniger-generell-als k'
 - k kompatibel-zu k' , bzw. überlappend
 - k inkompatibel-zu k' , bzw. disjunkt

Structural Knowledge

- Wissen über die Struktur der HC
- $Berge_1$ bezeichnet Bilder, keine Bücher

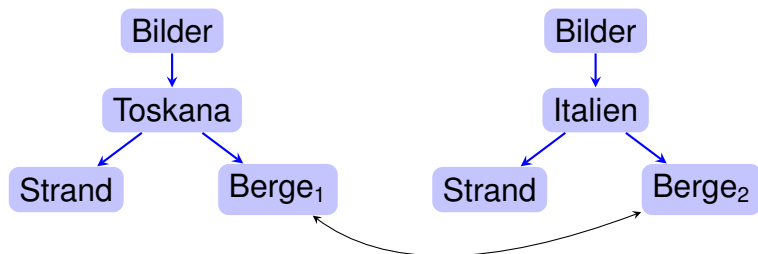
Lexical Knowledge

- Wissen über spezifische Wörter
- $Bild_a$ ist eine Darstellung
- $Bild_b$ ist ein Presseerzeugnis
- Auch: $Bild_a$ ist äquivalent zu $Darstellung_a$

Domain Knowledge

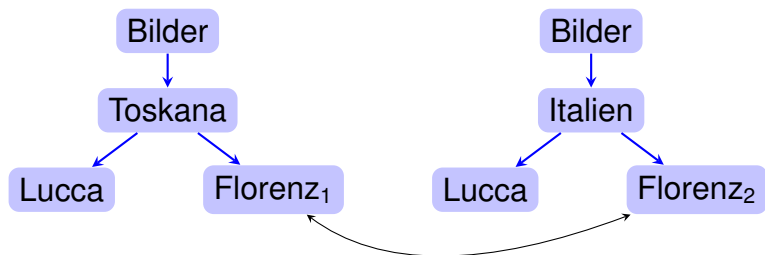
- Wissen über Relationen zwischen Bedeutungen von Wörtern
- Die Toskana ist Teil von Italien
- Florenz ist in Italien

Wissensarten Beispiel I



- Structural Knowledge
Berge₁ bezeichnet Bilder von Bergen in der Toskana
Berge₂ bezeichnet Bilder von Bergen in Italien
- Lexical Knowledge: Der gleiche Begriff
- Domain Knowledge
Toskana ist ein Teil von *Italien*
- Zusammen: Berge₁ ist weniger generell als Berge₂

Wissensarten Beispiel II



- Structural Knowledge
Florenz₁ bezeichnet Bilder von Florenz in der Toskana
Florenz₂ bezeichnet Bilder von Florenz in Italien
- Lexical Knowledge: Der gleiche Begriff
- Domain Knowledge
Florenz ist in der Toskana, Florenz ist in Italien
- Zusammen: Florenz₁ ist äquivalent zu Florenz₂

1 Einführung und Motivation

2 CTXMATCH

- Überblick

- Arbeitsweise

3 Gütekriterien

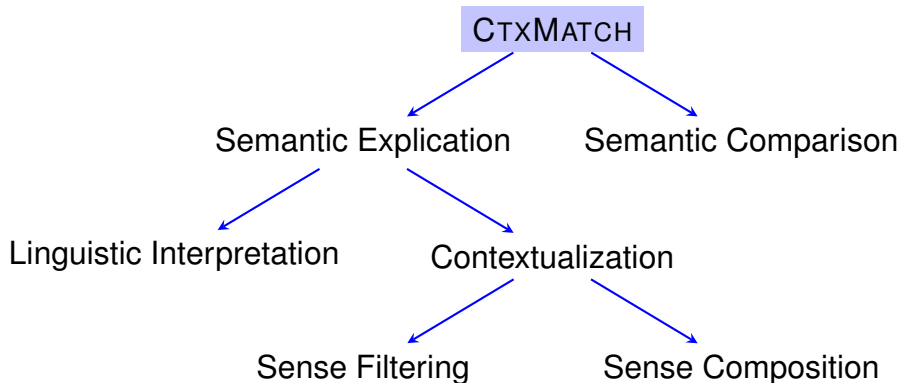
4 Zusammenfassung

Explication and Comparison

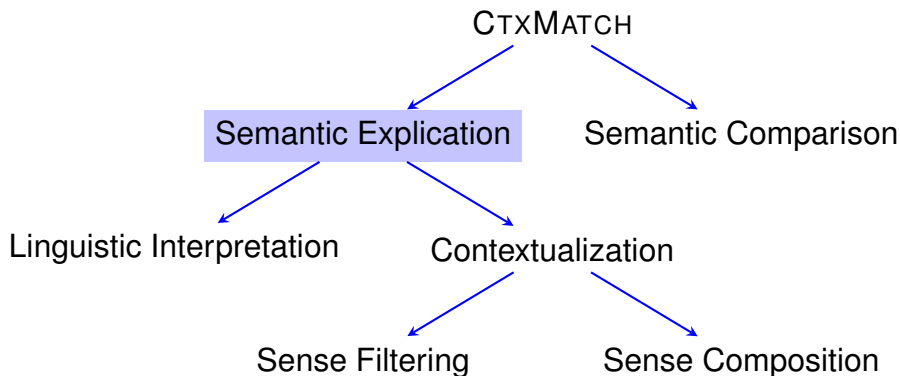
Semantic Explication Erstellt für jeden Knoten $k \in H$ eine Formel $w(k)$ in einer Logik, die diesen Knoten vollständig beschreibt

Semantic Comparison findet alle logischen Relationen zwischen $w(k)$ und $w(k')$

Struktur von CTXMATCH



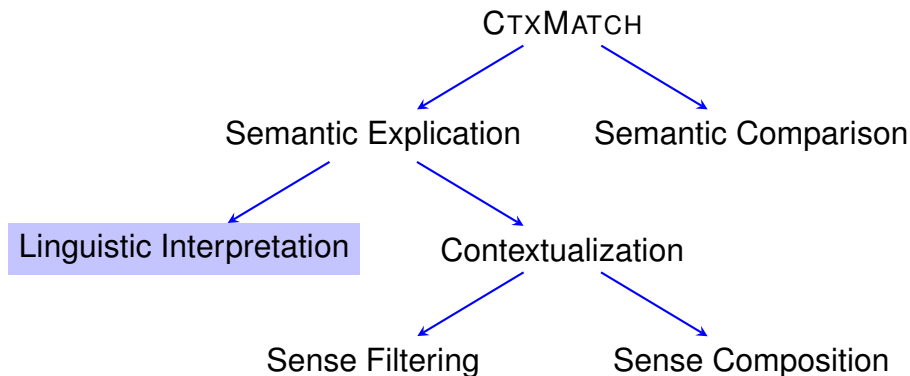
Struktur von CTXMATCH



Semantic Explication

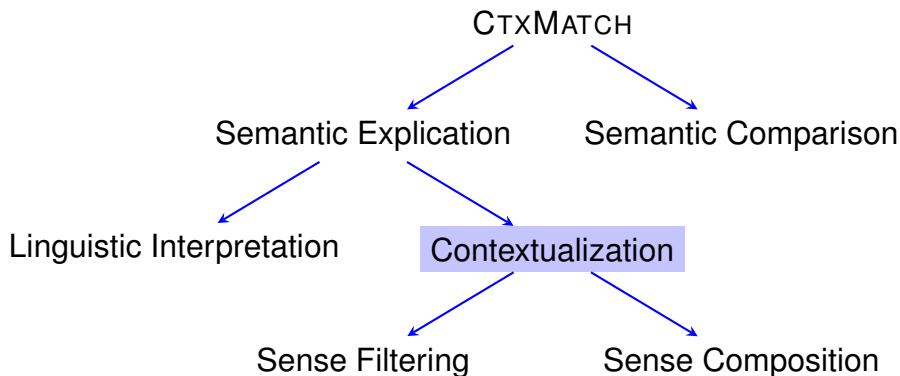
- *Semantic Explication* weist jedem Knoten eine logische Formel zu, die ihn vollständig beschreibt
- Hier: Aussagenlogik, Aussagensymbole sind Wortbedeutungen aus WORDNET
- Beispiel
 $\text{Florenz}_1 \Rightarrow \text{Florenz}_a \wedge \text{Toskana}_a \wedge \text{Bild}_a$
- Zwei Schritte
 - Linguistic Interpretation
 - Contextualization

Struktur von CTXMATCH



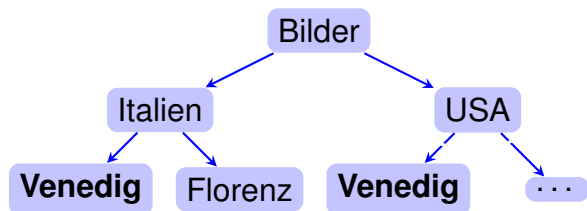
- Jede Beschriftung erhält eine Formel, die sie unabhängig von der Struktur beschreibt
- Beispiele
 - *Barock* \Rightarrow Barock_a
 - *Arizona* \Rightarrow Arizona_a \vee Arizona_b
 - *Bilder, Grafiken* \Rightarrow Bilder_a \vee Grafiken_a
 - *Klassische Musik*
 $\Rightarrow ((\text{Klassisch}_a \vee \dots) \wedge (\text{Musik}_a \vee \dots)) \vee \text{Klassische_Musik}_a$
 - ...

Struktur von CTXMATCH



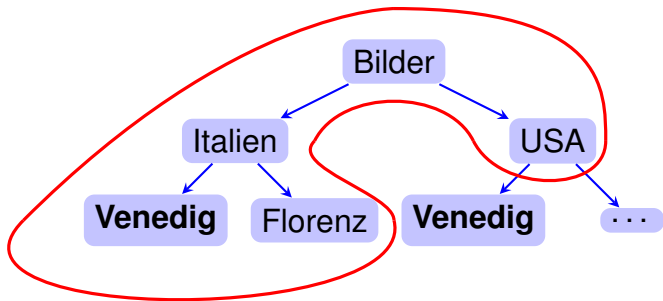
Contextualization

- Bedeutung durch die Position in der Hierarchie
- Hängt vom *Fokus* ab
- Fokus: Alle Elternknoten und deren direkte Kinderknoten

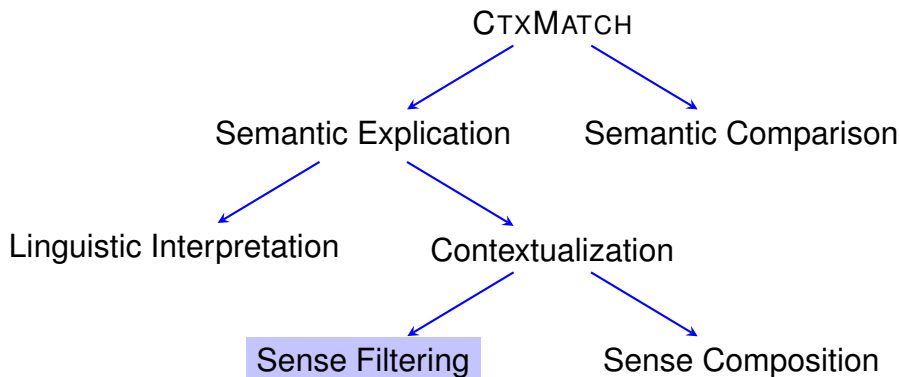


Contextualization

- Bedeutung durch die Position in der Hierarchie
- Hängt vom *Fokus* ab
- Fokus: Alle Elternknoten und deren direkte Kinderknoten



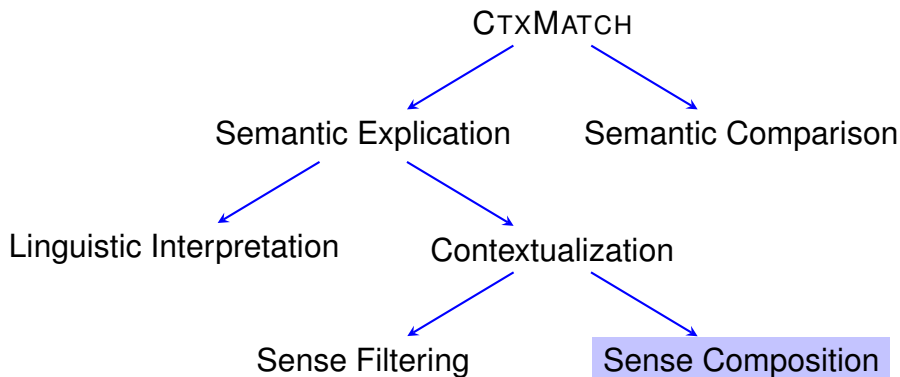
Struktur von CTXMATCH



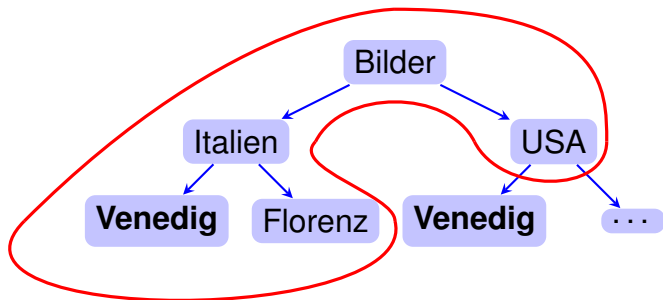
Sense Filtering

- Behält nur linguistische Interpretationen, die wir benötigen
- Zum Beispiel: Arizona₂ (die Schlange) kann entfernt werden, wenn in den Elternknoten der Sinn Wueste auftaucht, aber nichts mit Schlangen

Struktur von CTXMATCH

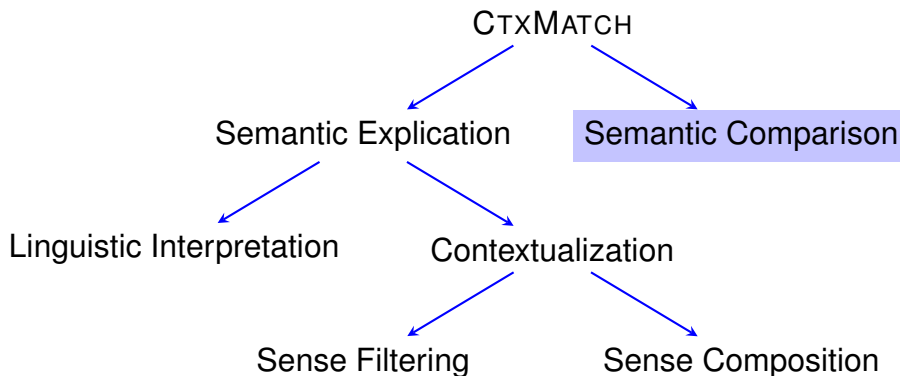


Sense Composition



- Menschen navigieren vom Wurzelknoten unter Ausschluss der anderen Optionen in jedem Knoten
- $\text{Venedig}_1 \Rightarrow \text{Venedig}_a \wedge \text{Italien}_a \wedge \neg \text{Florenz}_a \wedge \text{Bilder}_a \wedge \neg \text{USA}_a$

Struktur von CTXMATCH



Semantic Comparison

- Erstellt die Relationen zwischen den beiden HCs
- Verwendet dafür die Formeln, die im ersten Schritt für die Beschriftungen gefunden wurden
- Verwendet als Axiome eine *Background Theory*

Background Theory

- Die *Background Theory* enthält als Axiome alle Relationen zwischen allen Aussagensymbolen s_k (also $\text{Arizona}_a, \dots$)
- Relationen nach WORDNET
 - s_k synonym t_h
 $\Rightarrow s_k \equiv t_h$
 - s_k hyponym t_h
 $\Rightarrow s_k \rightarrow t_h$
 - s_k hypernym t_h
 $\Rightarrow s_k \leftarrow t_h$
 - s_k Gegenteil t_h
 $\Rightarrow \neg(s_k \wedge t_h)$

Übergang zur Logik

- Jede Beschriftung hat nun eine zugeordnete logische Formel
- Alle Relationen zwischen $w(k)$ und $w(k')$
- Relationen dazwischen sind nun nur noch eine Frage der logischen Deduktion
- Dank Informationen aus der *Background Theory*

Background Theory T

$T \models w(k_s) \rightarrow w(k_t)$	\Rightarrow	k_s genereller_als k_t
$T \models (w(k_t) \leftrightarrow w(k_s))$	\Rightarrow	k_s aequivalent_zu k_t
$T \models (w(k_s) \wedge w(k_t)) \rightarrow \perp$	\Rightarrow	k_s inkompatibel_zu k_t
$(k_s \wedge k_t)$ konsistent in T	\Rightarrow	k_s kompatibel_zu k_t

1 Einführung und Motivation

2 CTXMATCH

- Überblick
- Arbeitsweise

3 Gütekriterien

4 Zusammenfassung

Webverzeichnisse koordinieren

Koordinierung zwischen Google™ Directory und Yahoo!™ Directory.

Relation	Architecture		Medicine	
	Precision	Recall	Precision	Recall
Äquivalenz	0.71	0.10	0.78	0.13
Weniger-generell	0.85	0.49	0.88	0.46
Genereller	0.51	0.91	0.60	0.78

beinhaltet-Relation zwischen

- *Architecture / History / Medieval*
- *Architecture / History / Period and Styles / Gothic / Gargoyles*
- WORDNET: *Gothic* ist hyponym von *Medieval*

Reklassifizierung

Reklassifizierung aus privater Klassifizierung in ein Standard-Klassifizierungs-Framework (UNSPSC, Universal Standard Products and Services Classification)

	Stringbasiert		CTXMATCH	
Gesamt	194	100 %	194	100 %
Richtig	75	39 %	134	70 %
Falsch	91	50 %	16	8 %
Nicht klassifiziert	27	14 %	42	22 %

- CTXMATCH korrekter (70 % vs. 39 %), weniger Fehler (8 % vs. 50 %)

1 Einführung und Motivation

2 CTXMATCH

- Überblick
- Arbeitsweise

3 Gütekriterien

4 Zusammenfassung

Zusammenfassung

- Verwendet drei Typen von Wissen
- Verwendet die Semantik der Beschriftungen, nicht nur die Struktur
- Repräsentiert Knoten von HCs als logische Formeln
- Abbildung der HCs über logisches Schließen
- Problem wird verschoben: Vom Finden linguistischer/struktureller Ähnlichkeiten auf logisches Schließen

Fragen?