Ontology-based Multimedia Interpretation

Ralf Moeller

Hamburg Univ. of Technology

SSMS'08

Summer School on Multimedia Semantics '08

Acknowledgements

- Thanks to the members of the BOEMIE project (Bootstrapping Ontology Evolution for Multimedia IntEpretation, EU-FP6-IST)
- Members of my group
  - Sofia Espinosa Peraldi
  - Atila Kaya
  - Sylvia Melzer
  - Tobias Näth
  - Michael Wessel

Why media interpretation?

Multimedia documents

- E.g. web pages, videos
- Hierarchically structured documents
- Data vs. metadata

Content (Data)

- Raw data (modality-specific representation)
- Bag-of words model (vector of strings)
  - Easy for text documents
  - For images on web pages, take surrounding text to fill vectors
  - Manual annotations for image/video docs to form vectors
- Problem:
  - What are meaningful substructures (zones) in large docs?

Metadata

- Relational structure (author, resolution, …)
- Strings ("keywords")
  - Possibly taken from fixed repertoire or free text
  - Can be seen as part of relational structure
  - Can also be seen as part of bag-of-words model
  - Possibly organized into zones

Media retrieval...

- Standard:
  - Databases for metadata
    - Field-based retrieval
    - First-order logic and model checking (SQL)
  - Search engines supporting free text indexing (bag-of-words model for content)
    - Scoring functions (TF-IDF)
    - Querying (example-based)
      - Matching (nearest neighbor and friends)
    - Classification
      - Clustering techniques

… and content management

- Manipulate documents as objects (content plus metadata)
- Analyze content to derive topics: Which clusters show up?
- Find subsets of all documents in a repository that might be of interest for a customer to view or to buy

Desiderata?

- Bag of words model does not take the syntactic structure of text into account (not even shallow syntactic relations are considered)
- Bag-of-words model hard to implement for images and video
  - Surrounding text not always available
  - Manual annotation “expensive”
    - Depends on the view/mood of the annotator
    - Usually not (re)done in the light of subsequent events

Media Interpretation (MI)

Possible interpretations

- "Two workers empty garbage containers into a garbage truck"
- "A mailman distributes mail"

What is a possible high-level annotation for this image?

Interpretations = Metadata (1)

$$
\begin{array}{rcl}
mailman_1 & : & Mailman \\
bicycle_1 & : & Bicycle \\
mail\_deliv_1 & : & MailDeliv \\
(mail\_deliv_1, mailman_1) & : & hasPart \\
(mail\_deliv_1, bicycle_1) & : & hasPart \\
(mail\_deliv_1, url\_1) & : & hasURL \\
(mailman_1, url\_2) & : & hasURL \\
(bicycle_1, url\_3) & : & hasURL \\
(url\_1) & : & =\texttt{"http://www.img.de/image-1.jpg"} \\
(url\_2) & : & =\texttt{"http://www.img.de/image-1.jpg\#(200,400)/} \\
(url\_3) & : & =\texttt{"http://www.img.de/image-1.jpg\#(100,400)/}
\end{array}
$$

Interpretations = Metadata (2)

$$\begin{array}{rcl} & \ldots & \\ garbageman_1 & : & Garbageman \\ garbageman_2 & : & Garbageman \\ garbagetruck_1 & : & Garbage\_Truck \\ gc_1 & : & Garbage\_Collection \\ (gc_1, garbageman_1) & : & hasPart \\ (gc_1, garbageman_2) & : & hasPart \\ (gc_1, garbagetruck_1) & : & hasPart \\ (gc_1, url\_4) & : & hasURL \\ & \ldots & \end{array}$$

Media access (MA)

- Queries:

$$\begin{aligned} ImageQuery_1 &:= \{(X,Y) \mid & MailDeliv(X), Bicycle(Y), hasPart(X,Y)\} \\ URLQuery_1 &:= \{(X, value(X)) \mid & hasURL(mail\_deliv_1, X)\} \\ URLQuery_2 &:= \{(X, value(X)) \mid & hasURL(bicycle_1, X)\} \end{aligned}$$

- Results :
  - { $(mail\_deliv_1, bicycle_1)$ }
  - { $(url_1,$ "http://www.img.de/image-1.jpg") }
  - { $(url_3,$ "http://www.img.de/image-1.jpg#(100,400)/(150/500)) }

- Exploit ontologies for querying:

$$Mailman \sqsubseteq Postal\_Employee$$

$$Mailman \equiv Postman$$

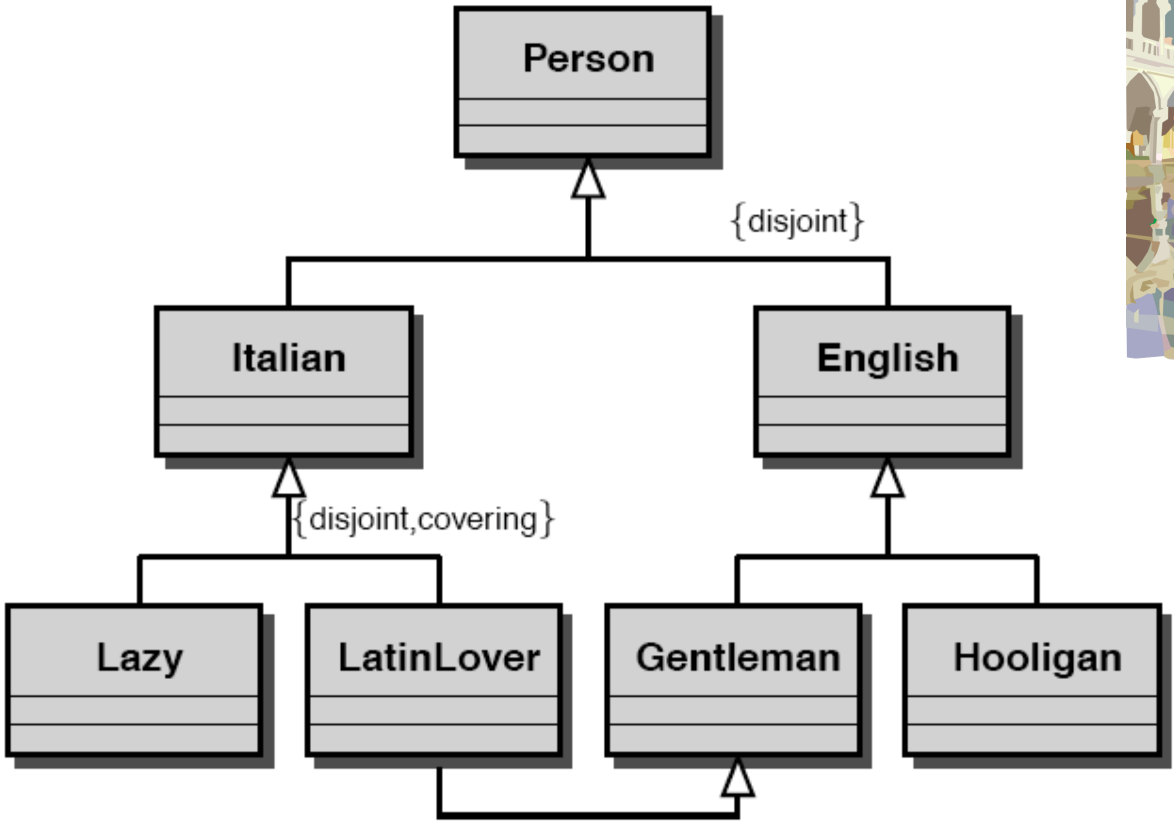
Logic-based content management?

- Ontologies are based on first-order logic
  - Description logic
    - Tbox (terminological knowledge)
    - Abox (assertional knowledge)
  - Horn logic
    - Logic programs ("Rules")
- First-order logic combined with
  - Fuzzy-set theory
  - Probability theory

Ontologies & Reasoning

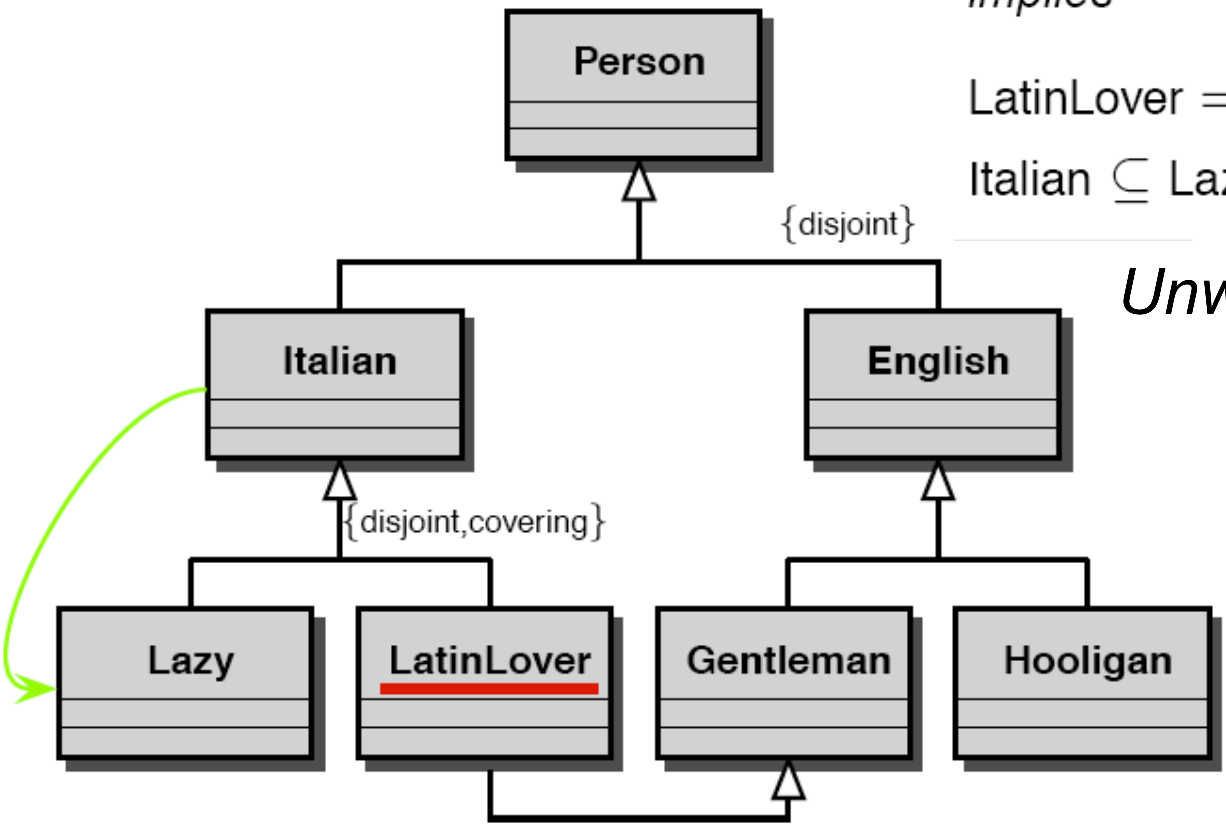
- Ontology = (Tbox, Abox)
- Tbox (= set of axioms)
  - Find implictly stated axioms
  - Find unwanted implicit axioms/facts
- Abox (= set of assertions)
  - Find implicit assertions or combinations of implicit assertions
  - Find unwanted implicit assertions

Where are the Latin lovers?



Example from E. Franconi (Univ. Bolzano)

Modeling error revealed



implies

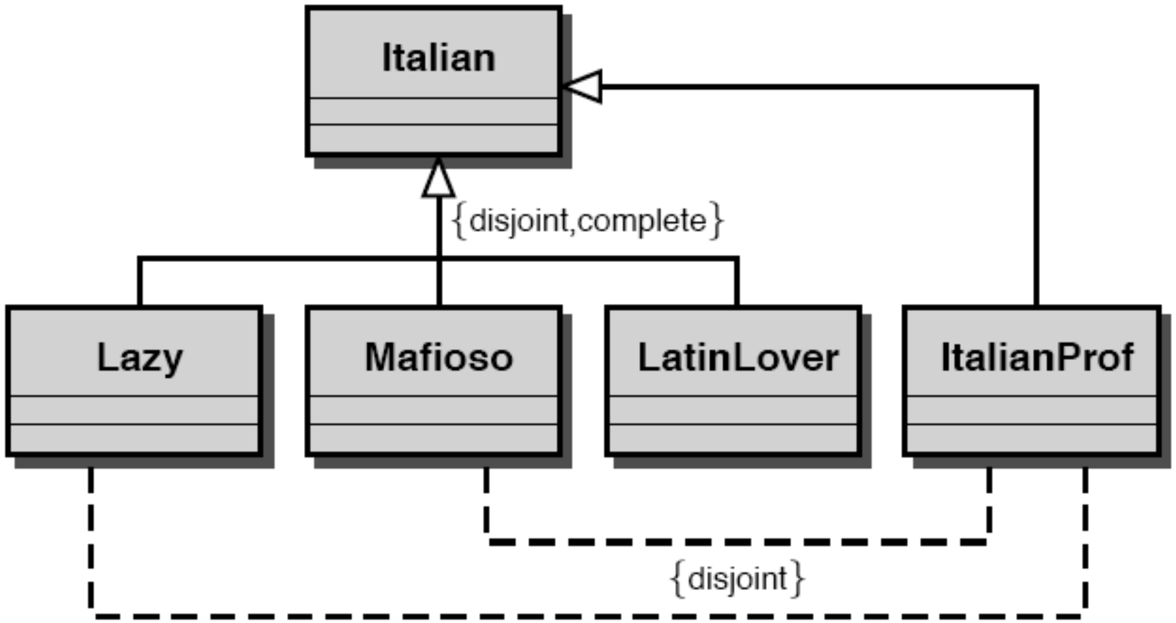
Inconsistency

$LatinLover = \emptyset$

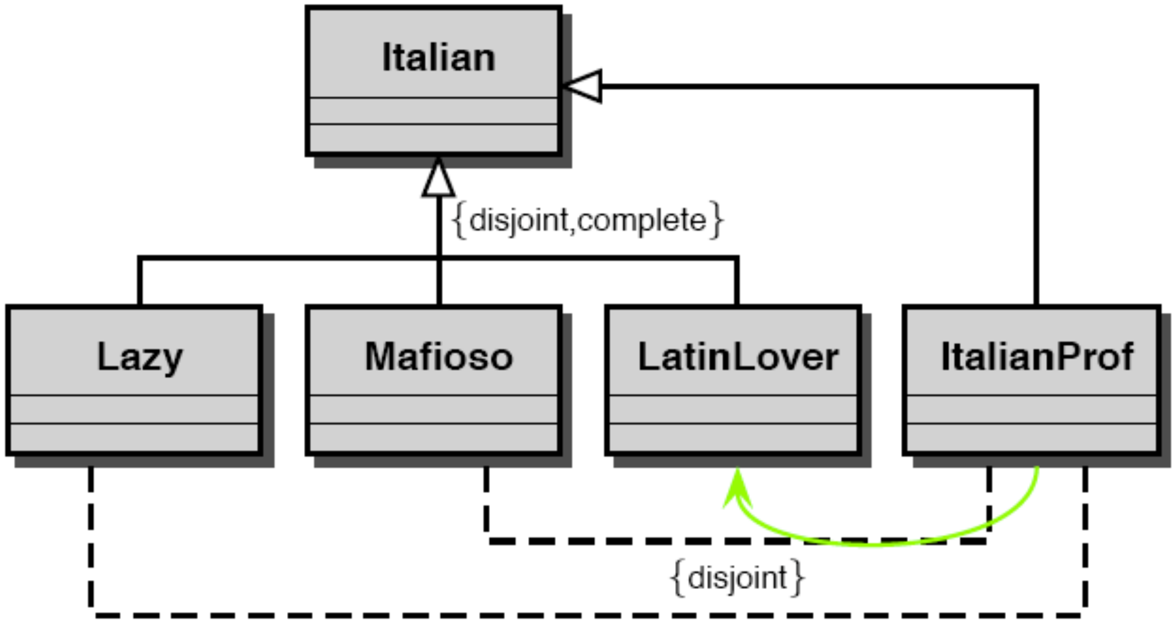
$Italian \subseteq Lazy$ – $Italian \equiv Lazy$

Unwanted subsumption

Oh those Italian Professors



Modeling error – unfortunately



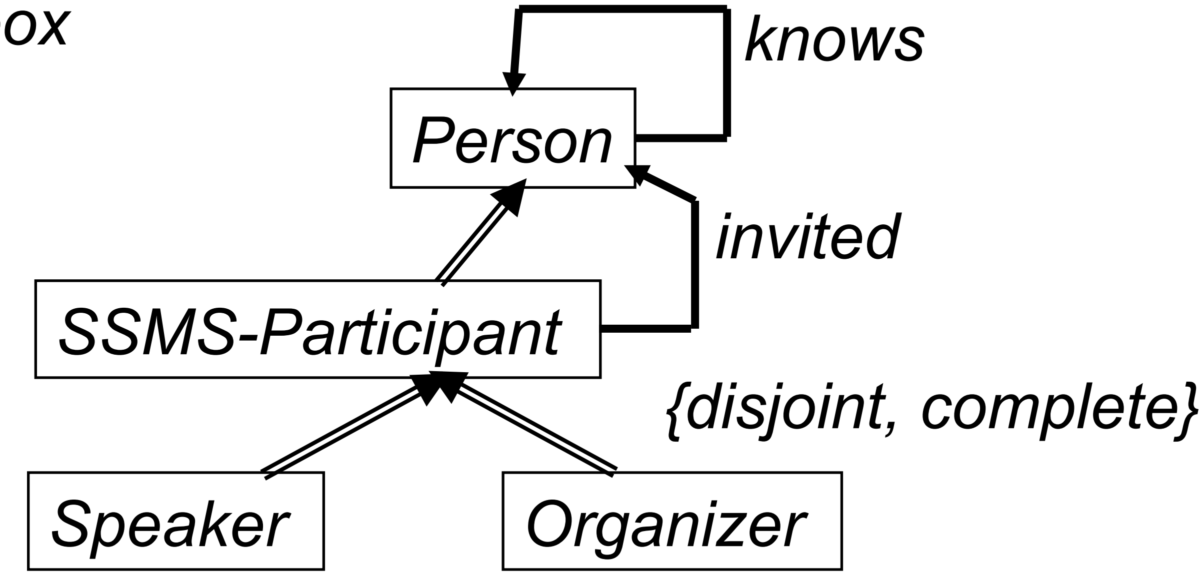
implies

ItalianProf $\subseteq$ LatinLover

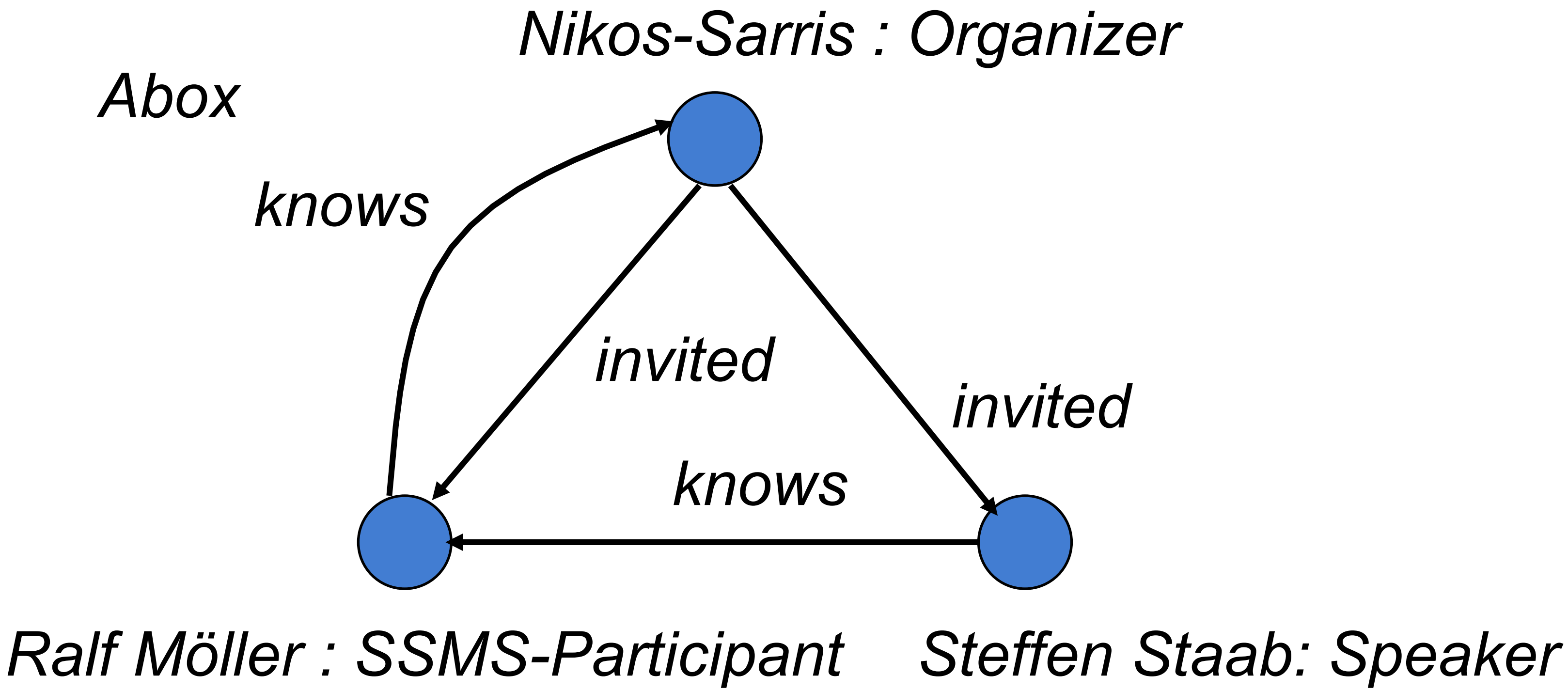
Unwanted subsumption?

Reasoning for query answering

Tbox



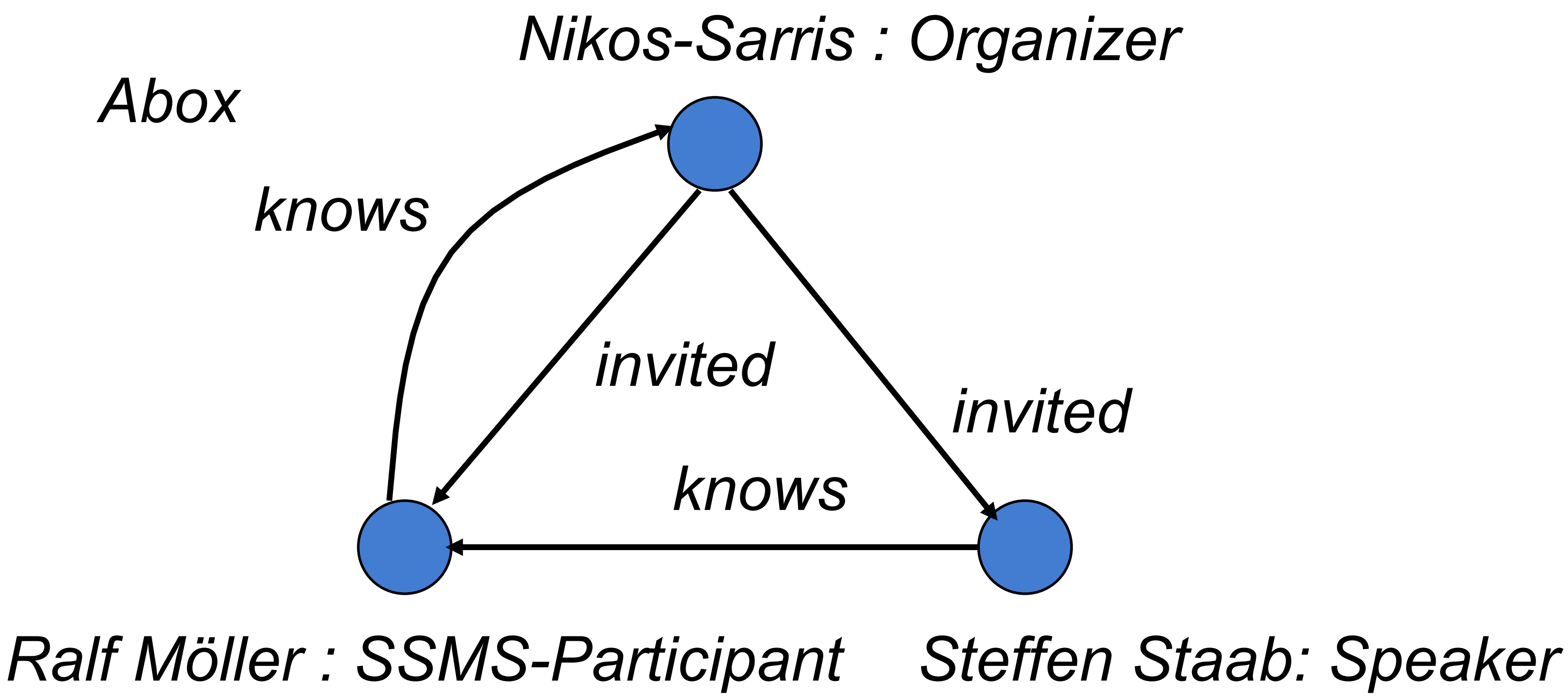
Example: Data descriptions



Query

Bad presentation? Who is responsible?
„I am looking for a person who invited a speaker who knows an organizer.“

Example: Data descriptions



Query { (X) | Person(X), invited(X, Y), Speaker(Y), knows(Y, Z), Organizer(Z) }

Why is all this relevant for MMI?

- Ontology development (deduction)
- Query answering (deduction)
- Metadata derivation
  by interpretation (fomalized as abduction, see below)

How to derive metadata?

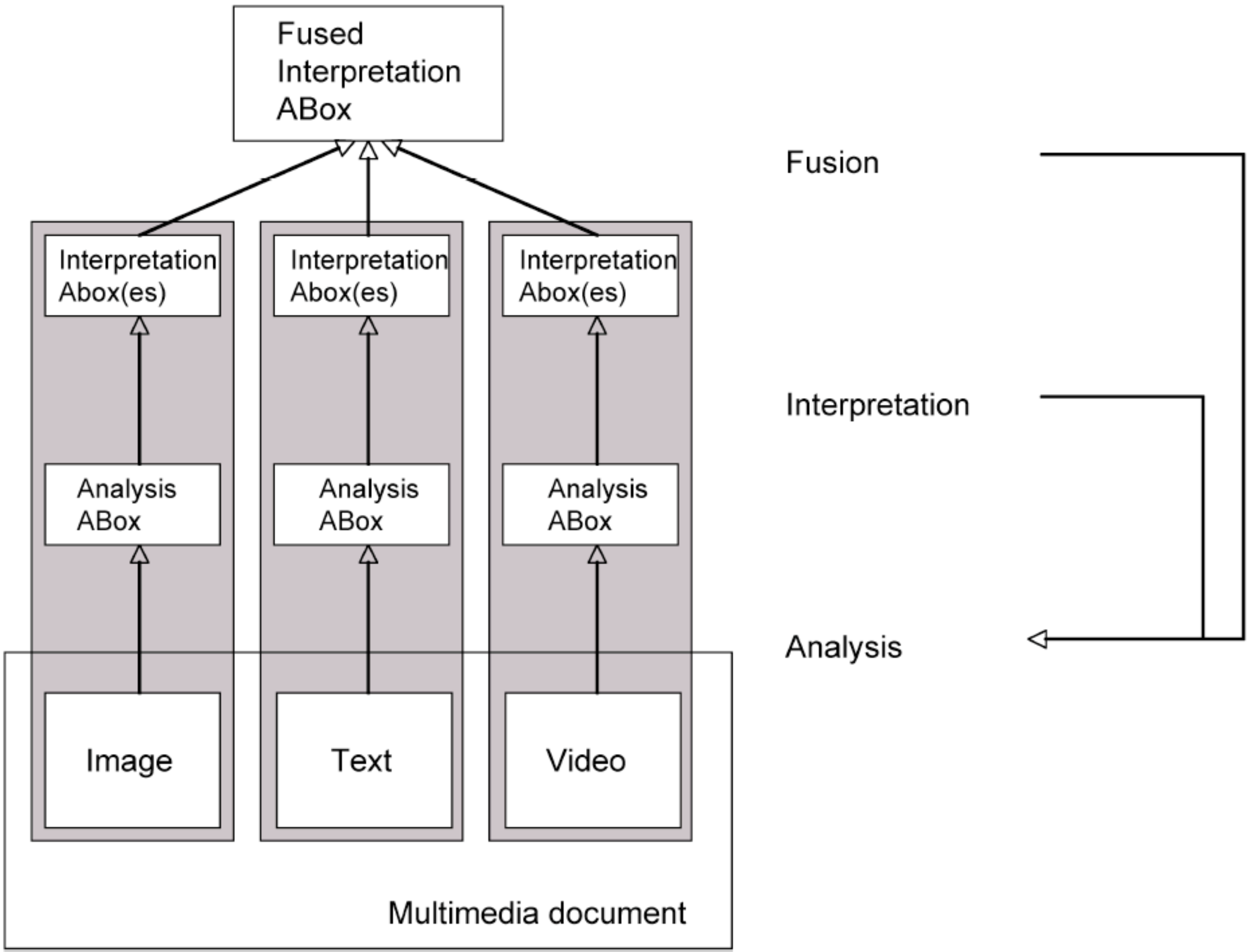

Fused
Interpretation
ABox
Fusion
Interpretation
Abox(es)
Interpretation
Abox(es)
Interpretation
Abox(es)
Interpretation
Analysis
ABox
Analysis
ABox
Analysis
ABox
Analysis
Image
Text
Video
Multimedia document

Image analysis

- Assumption: Result of image analysis represented as an Abox

$$
\begin{aligned}
pole_1 &: Pole \\
human_1 &: Human \\
bar_1 &: Bar \\
(bar_1, human_1) &: near
\end{aligned}
$$

Querying for a pole vault?

- Image will not be found
- Reason:
  - Pole vaults are not directly visible
  - There is no Abox individual for a pole vault
- Idea:
  - Pole vaults are "constructed" to explain the spatial configuration of a person, a bar, and a pole

Interpretation as explanation

- Given: an (explicit) assertion α
- Find a derivation which explains why α is true
- Basis
  - Domain knowledge (Tbox, Abox)
  - Accepted observations (Abox)
  - Explanation knowledge (formulated as logic program)
- Explanation = set of assertions

Explanation via abduction

- Compute some set of assertions Δ such that

$$\Sigma \cup \Delta \models \Gamma$$

- The knowledge base $\Sigma = (\mathcal{T}, \mathcal{A})$
- The set of observables:

Γ is a given set of low-level assertions

- Explanation knowledge determines |=
  - Set of rules (logic program)

Rules: an example

$$near(Y,Z) \leftarrow Pole\_Vault(X), hasPart(X,Y), Bar(Y),$$
$$hasPart(X,W), Pole(W), hasParticipant(X,Z), Jumper(Z)$$
$$near(Y,Z) \leftarrow High\_Jump(X), hasPart(X,Y), Bar(Y),$$
$$hasParticipant(X,Z), Jumper(Z)$$

In our setting

- Image analysis leads to an Abox

$$\begin{aligned} pole_1 &: Pole \\ human_1 &: Human \\ bar_1 &: Bar \\ (bar_1, human_1) &: near \end{aligned}$$

- Assumption: last assertion to be explained, rest taken for granted

$$\Sigma \cup \Gamma_1 \cup \Delta \models \Gamma_2$$

- The Abox is split into parts Γ_1 and Γ_2

Side conditions

- Minimality
- Consilience

Example Tbox

$$
\begin{aligned}
Jumper &\sqsubseteq Human \\
Pole &\sqsubseteq Sports\_Equipment \\
Bar &\sqsubseteq Sports\_Equipment \\
Pole \sqcap Bar &\sqsubseteq \bot \\
Pole \sqcap Jumper &\sqsubseteq \bot \\
Jumper \sqcap Bar &\sqsubseteq \bot \\
Jumping\_Event &\sqsubseteq \exists_{\leq 1} hasParticipant.Jumper \\
Pole\_Vault &\sqsubseteq Jumping\_Event \sqcap \exists hasPart.Pole \sqcap \exists hasPart.Bar \\
High\_Jump &\sqsubseteq Jumping\_Event \sqcap \exists hasPart.Bar \\
near(Y,Z) &\leftarrow Pole\_Vault(X), hasPart(X,Y), Bar(Y), \\
&\quad hasPart(X,W), Pole(W), hasParticipant(X,Z), Jumper(Z) \\
near(Y,Z) &\leftarrow High\_Jump(X), hasPart(X,Y), Bar(Y), \\
&\quad hasParticipant(X,Z), Jumper(Z)
\end{aligned}
$$

Abduction as query answering

- Abduction equation

$$\Sigma \cup \Gamma_1 \cup \Delta \models \Gamma_2$$

- Entailment of assertions as answering boolean queries w.r.t. Tboxes and Aboxes

$$Q_1 := \{() \mid near(bar_1, human_1)\}$$

Abductive query answering

- The answer to Q_1 should be true!
- What must be added to the Abox to ensure this?
- Idea: Look at the rules
  - Assumption: The rules define the set of abducibles
  - Apply the rules in a backward-chaining way

Backward-chaining

- $$near(Y,Z) \leftarrow Pole\_Vault(X), hasPart(X,Y), Bar(Y),$$
  $$hasPart(X,W), Pole(W), hasParticipant(X,Z), Jumper(Z)$$
  $$near(Y,Z) \leftarrow High\_Jump(X), hasPart(X,Y), Bar(Y),$$
  $$hasParticipant(X,Z), Jumper(Z)$$
- $$Q_1 := \{() \mid near(bar_1, human_1)\}$$
- Determination of a variable substitution (aka variable binding)
  - Match rule head with query
  - Introduce new individuals for yet unbound variables or take individual mentioned in the Abox
- Nested calls due to logic programs (no recursion)

$$\{\{atom_1, atom_2, \ldots atom_k\}\}$$

Example

- $\Delta_1 = \{new\_ind_1 : Pole\_Vault, (new\_ind_1, bar_1) : hasPart, (new\_ind_1, new\_ind_2) : hasPart, new\_ind_2 : Pole, (new\_ind_1, human_1) : hasParticipant, human_1 : Jumper\}$
- $\Delta_2 = \{new\_ind_1 : Pole\_Vault, (new\_ind_1, bar_1) : hasPart, (new\_ind_1, pole_1) : hasPart,$
  $(new\_ind_1, human_1) : hasParticipant, human_1 : Jumper\}$
- $\Delta_3 = \{new\_ind_1 : High\_Jump, (new\_ind_1, bar_1) : hasPart, (new\_ind_1, human_1) : hasParticipant,$
  $human_1 : Jumper\}$

Why is reasoning important?

- Assumption: Explicit assertions need to be explained (if possible)
- Avoid the generation of assertions during backward chaining
- Reasoning required to determine which assertions already hold

Forward chaining

- If there is an object *x* with a part *y* with *y* being *near* another object *z*, then the whole thing *x* is *near z*

- *Thing(X), hasPart(Y), near(Y, Z)*
  → *near(X,Z)*

Preference measures

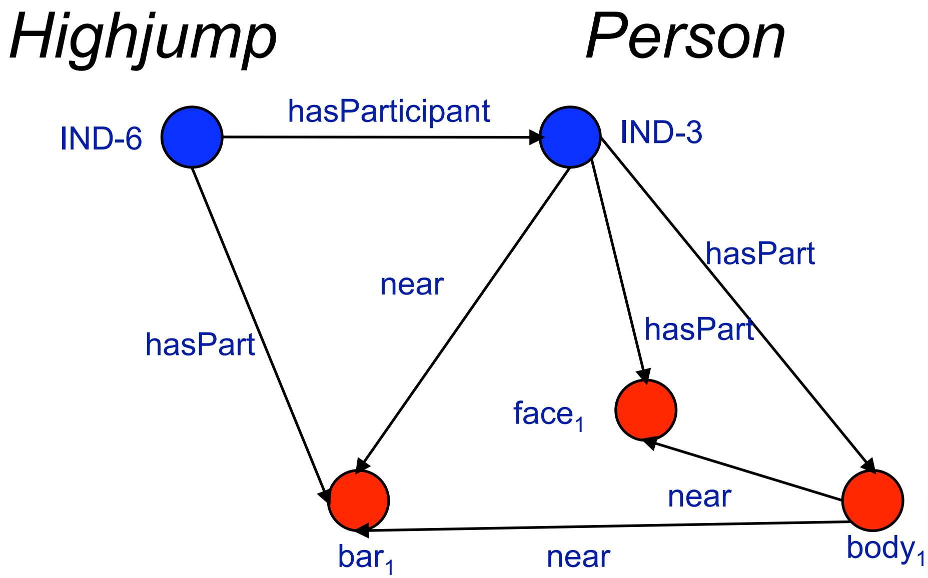
- When more than one explanation is found, the criteria for selecting the best explanation is:
  - Simplicity
  - Consilience

$$S(\Delta) := S_f(\Delta) - S_h(\Delta)$$

$$S_f(\Delta) := |\{i | i \in assertions(\Delta) \text{ and } i \in assertions(\Gamma_1)\}|$$
$$S_h(\Delta) := |\{i | i \in newAssertions(\Delta)\}|$$

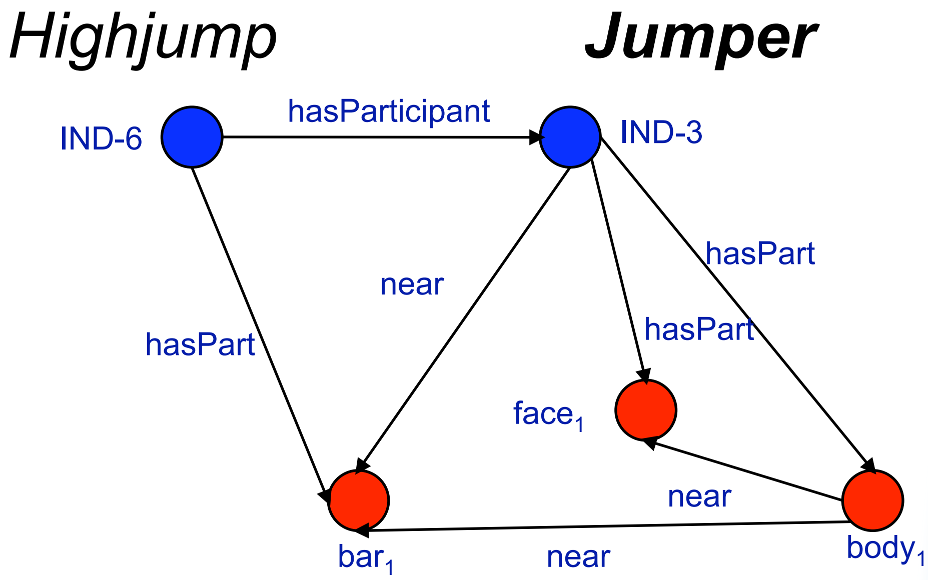
Example 2: Image modality

- Abduction constructs relational structures
- Tbox causes new knowledge to be inferred



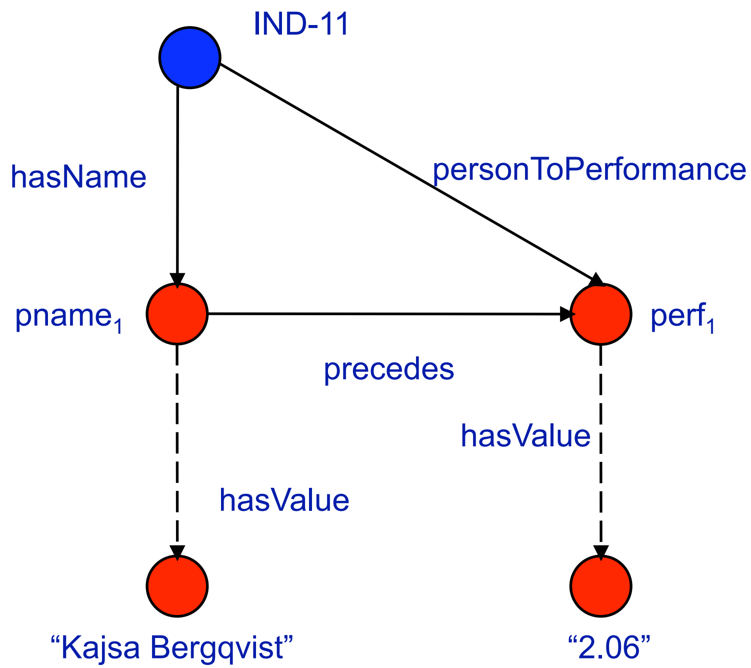
Example 2: Image modality

- Abduction constructs relational structures
- Tbox causes new knowledge to be inferred



Example 3: Text modality

Kajsa Bergqvist clears 2:06 in Eberstadt

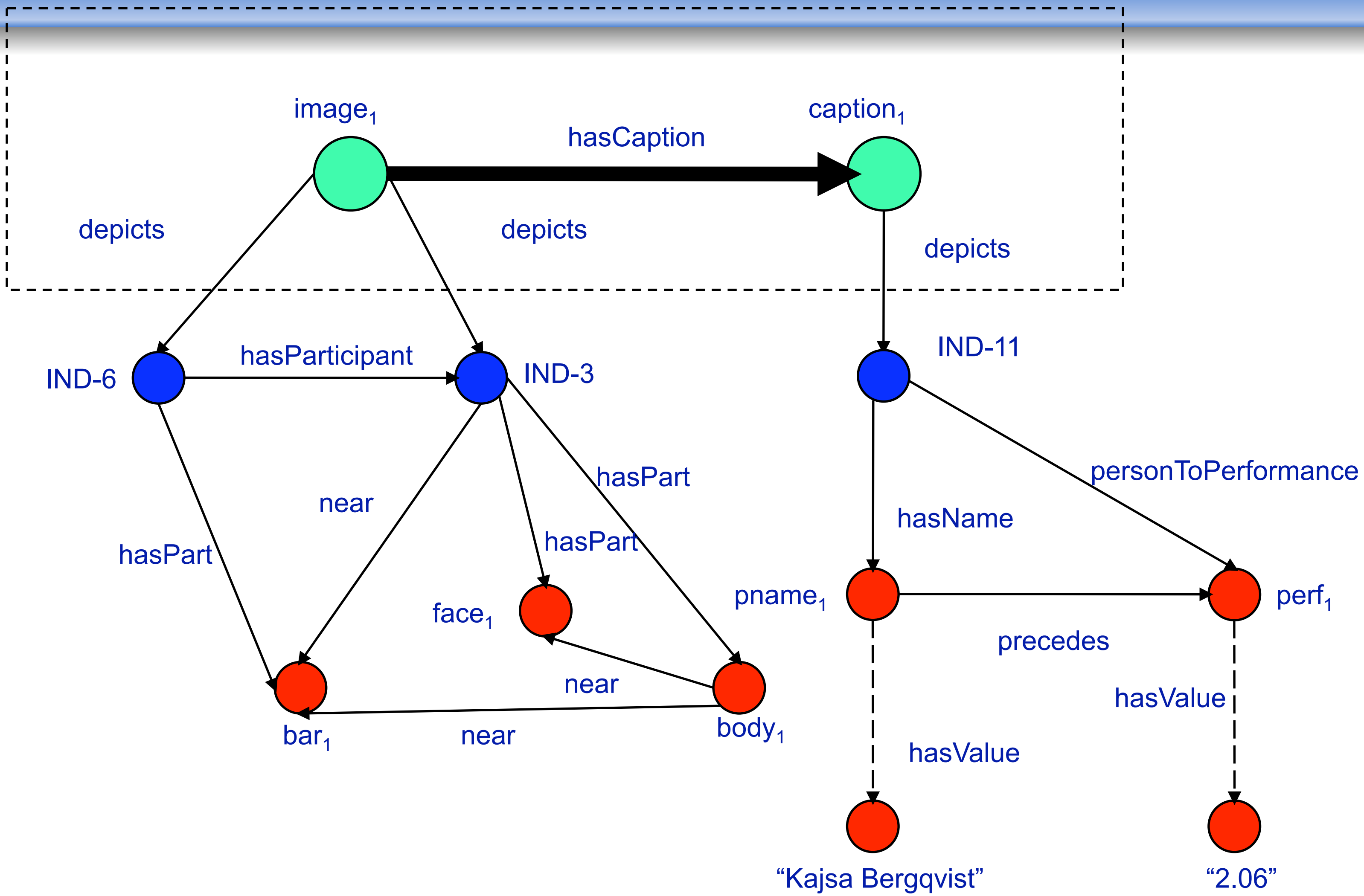


Fusion

- Goals:
  - Desambiguation
  - Rule out possible interpretations
- Information accumulation (for better query answering)

Kajsa Bergqvist clears 2:06 in Eberstadt

Explicitly represent the document structure

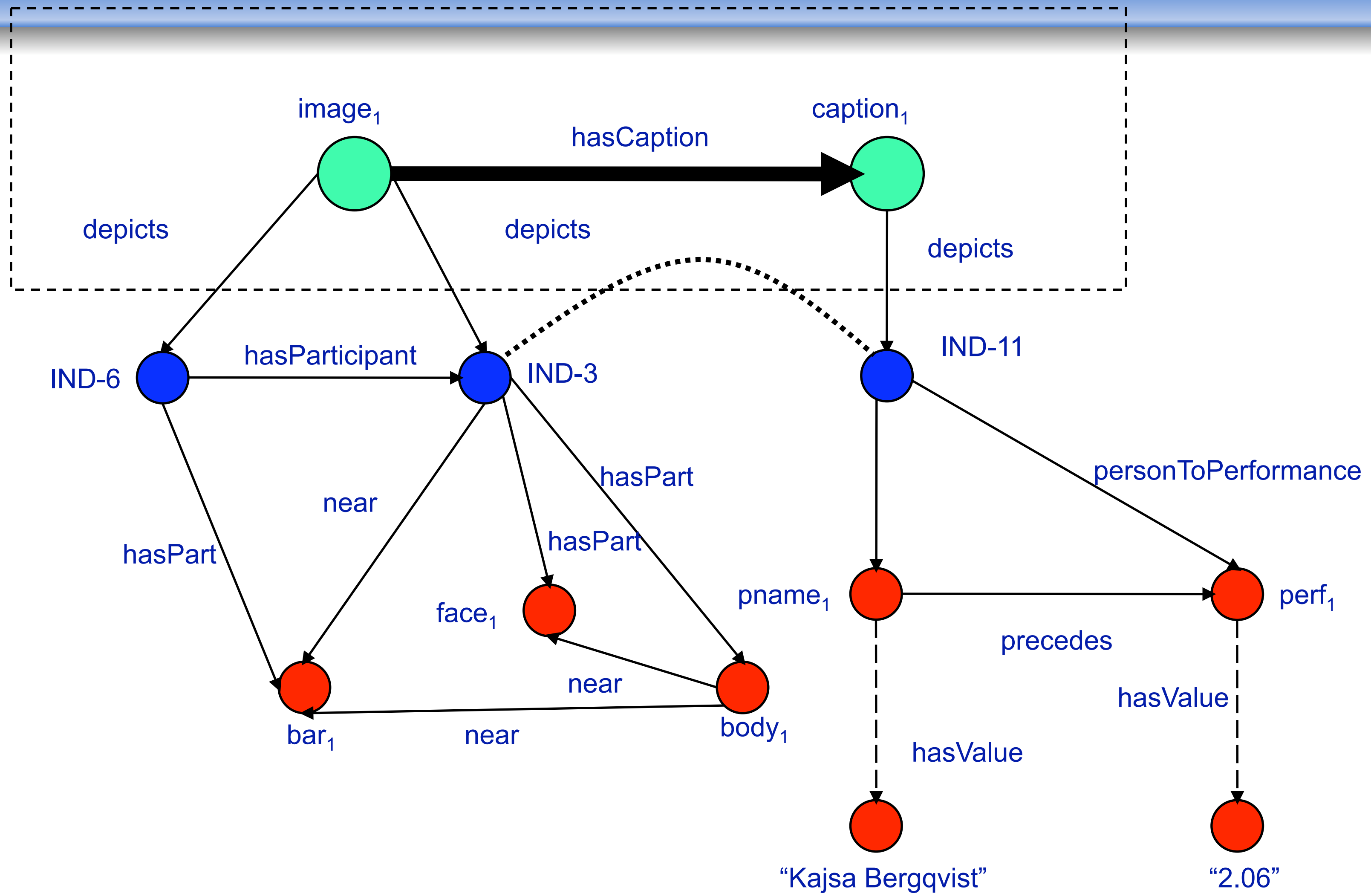


Exploit the document structure

- Abduction is used to find explan
a
tions for the relations between the multimedia objects
- We assume the following rule

```
hasCaption(X,A) :-
      Image(X),
      depicts(X,Y),
      Caption(A),
      depicts(A,B),
      same-as(Y,B)
```

Structure identification



Fusion for multimodal docs

Role of MI for retrieval

- Boolean Abox query answering with reasoning
- Augmentation of bag-of-words model with interpretation assertions (their consequences)

Event recognition in video

- Can-do assumptions
  - Object detection
  - Phases detection
    - What holds for an object in a maximal time interval?

Temporal propositions (1)

- Syntax: Let *Preds* be a set of predicate names.

A *temporal proposition* is a syntactic structure of the following form:

$$P_{[t_1,t_2]}(ind_1, \ldots, ind_n)$$

where t_i denotes an element of a linear temporal structure $\Theta \subseteq \mathbb{N}$, ind_i with $i \in \{1, \ldots, n\}$ denotes an individual, and $P \in Preds$.

- Example

$$\begin{aligned} accelerate\_horizontally_{[219,224]} \quad & (moving\_object_1) \\ vertical\_upward\_movement_{[224,226]} \quad & (moving\_object_1) \\ turn_{[226,228]} \quad & (moving\_object_1) \\ vertical\_downward\_movement_{[228,230]} \quad & (moving\_object_1) \end{aligned}$$

TPs as part of an Abox

$$
\begin{aligned}
accelerate\_horizontally_{[219,224]} & \quad (moving\_object_1) \\
vertical\_upward\_movement_{[224,226]} & \quad (moving\_object_1) \\
turn_{[226,228]} & \quad (moving\_object_1) \\
vertical\_downward\_movement_{[228,230]} & \quad (moving\_object_1) \\
moving\_object_1 : & \quad Jumper \\
event_1 : & \quad High\_Jump \\
(event_1, moving\_object_1) : & \quad hasPart
\end{aligned}
$$

Temporal propositions (2)

- Let $\Theta \subseteq \mathbb{N}$ be a linear temporal structure.
- A temporal interpretation $\mathcal{I}_T$ is a tuple $(\Delta, \cdot^{\mathcal{I}}, \Theta, \Im)$ such that $\Im$ is an injective mapping from the temporal structure Θ to a set of standard Tarskian interpretation functions.
- A temporal interpretation $\mathcal{I}_T = (\Delta, \cdot^{\mathcal{I}}, \Theta, \Im,)$ satisfies a GCI or an ABox assertion if the standard part $(\Delta, \cdot^{\mathcal{I}})$ satisfies the GCI or the ABox assertion.
- A temporal interpretation $\mathcal{I}_T$ satisfies a temporal proposition $P_{[t_1,t_2]}(ind_1, \ldots, ind_n)$ if the predicate is true for all time points in the non-empty interval $[t_1, t_2]$.

Temporal propositions (3)

- We assume that temporal propositions are durative, i.e., the proposition holds for all non-empty subintervals.
  More formally:

- $$\mathcal{I}_T \models P_{[t_1,t_2]}(ind_1, \ldots ind_n)$$

  if for all $\theta \in \Theta$ it holds that if $t_1 \leq \theta \leq t_2$, $|\Theta| > 1$, then $(ind_1^{\Im(\theta)}, \ldots, ind_n^{\Im(\theta)}) \in P^{\Im(\theta)}$.

- As usual, a temporal interpretation that satisfies a temporal proposition is called a <u>*temporal model*</u> for this term. A temporal interpretation which satisfies a GCI or an ABox assertion is called a (temporal) model for the GCI or ABox assertion, respectively.

Rules with time intervals (1)

- Syntax

$$
\begin{aligned}
P_{[T_0,T_1]}(X_1,\ldots,X_{n_1}) \quad \leftarrow \quad & Q_{1\,[T_2,T_3]}(Y_{1,1},\ldots,Y_{1,m_1}),\\
& \ldots\\
& Q_{k\,[T_{2k},T_{2k+1}]}(Y_{k,1},\ldots,Y_{k,m_k}),\\
& A_1(Z_1),\\
& \ldots\\
& A_l(Z_l),\\
& R_1(W_1,W2),\\
& \ldots\\
& R_h(W_{2h-1},W_{2h}),
\end{aligned}
$$

  Note that the variables $X_i, Y_{i,j}, Z_i$, and W_i are not necessarily disjoint.

- Rules must be non-recursive (with the obvious definition)

Rules with time intervals (2)

- Example

$$
\begin{aligned}
HJ\_Occurs_{[T_1,T_2]}(X) \quad \leftarrow \quad & High\_Jump(X), \\
& HJ\_InJumpPhase(X), \\
& hasPart(X,Y), \\
& Jumper(Y), \\
& vertical\_upward\_movement_{[T_1,T_3]}(Y), \\
& turn_{[T_3,T_4]}(Y), \\
& vertical\_downward\_movement_{[T_4,T_2]}(Y).
\end{aligned}
$$

Queries w.r.t. TPs (1)

- Syntax

$$
\begin{array}{lll}
\{(X_1, \ldots, X_n)_{[T_1,T_2]} & | & Q_{1[T_2,T_3]}(Y_{1,1}, \ldots, Y_{1,m_1}), \\
 & & \ldots \\
 & & Q_{k[T_{2k},T_{2k+1}]}(Y_{k,1}, \ldots, Y_{k,m_k}), \\
 & & A_1(Z_1), \\
 & & \ldots \\
 & & A_l(Z_l), \\
 & & R_1(W_1, W2), \\
 & & \ldots \\
 & & R_h(W_{2h-1}, W_{2h})\}
\end{array}
$$

- Grounded conjunctive queries
- Unfold atoms w.r.t. rules in queries
- Results in unions of grounded conjunctive queries

Queries w.r.t. TPs (2)

- Example

$$\{(X)_{[T_1,T_2]} \mid HJ\_Occurs_{[T_1,T_2]}(X), hasPart(X,Y), Famous(Y)\}$$

Answers to queries

A tuple $(ind_1, \ldots, ind_n)_{[t_1,t_2]}$ is a potential solution of a grounded unfolded temporal conjunctive query if the variable substitution $[X_1 \leftarrow ind_1, \ldots, X_n \leftarrow ind_n, T_1 \leftarrow t_1, T_2 \leftarrow t_2]$ can be extended with additional assignments for all other variables in the body such that the resulting query atoms after applying the substitution are satisfied in all temporal models of the ontology (TBox and ABox).

The result set for a temporal query comprises all tuples $(ind_1, \ldots, ind_n)_{[(t_{1_{min}}, t_{1_{max}}),(t_{2_{min}}, t_{2_{max}})]}$ such that there exists no other potential solution $(ind_1, \ldots, ind_n)_{[t_1,t_2]}$ with $t_1 < t_{1_{min}}$ or $t_1 > t_{1_{max}}$ or $t_2 < t_{2_{min}}$ or $t_2 > t_{2_{max}}$.

Algorithms for query answering

- Standard query atoms:
  - As for GCQs introduced before
- Temporal query atoms:
  - Matching of query atoms and temporal propositions in the Abox
    - Match([201, 233], [t_1, t_2])
  - Gives bindings for time variables as well as object variables
  - Manage temporal constraints
    - E.g., [t_1, t_2] gives rise to t_1,< t_2
    - Equality constraints for bindings
    - Satisfiability checking for constraints
    - Compuation of minimal/maximal values for time variables

Abduction with time variables

If a high-jump event is expected but the query for the high-jump event (see above) returns *false*, then abduction can be used to determine what has to be added to the interpretation ABox. For instance, the temporal proposition

$$accelerate\_horizontally_{[219,224]}(moving\_object_1)$$

might probably be missing, and will be added by abduction such that the answer will be *true* and the high jump event is "explained".

Repair

It might also be the case that the image sequence analysis determined a mutilated partial basic event such as

$$accelerate\_horizontally_{[219,223]}(moving\_object_1)$$

instead. In this case, abduction would just add the proposition as in the case before. However, in this case we prefer that a near-miss is recognized, and believe that a "repair" operation for the assertion in the analysis ABox should be proposed.

Modality fusion again

- Answering a query w.r.t. a single modality might not return a result
- Idea: combine interpretation Aboxes of multiple modalities

Example

- From image/text interpretation:

$$\begin{aligned} object_1 &: Event \\ object_2 &: Horizontal\_Bar \\ object_3 &: Athlete \\ object_3 &: Famous \\ (object_1, object_2) &: precedes \\ (object_1, object_3) &: precedes \end{aligned}$$

$$\begin{aligned} object_1 &: High\_Jump \\ (object_1, object_2) &: hasPart \\ (object_1, object_3) &: hasPart \\ object_1 &: HJ\_InJumpPhase \end{aligned}$$

- From video interpretation:

$$\begin{aligned} accelerate\_horizontally_{[219,224]} &\quad (moving\_object_1) \\ vertical\_upward\_movement_{[224,226]} &\quad (moving\_object_1) \\ turn_{[226,228]} &\quad (moving\_object_1) \\ vertical\_downward\_movement_{[228,230]} &\quad (moving\_object_1) \\ moving\_object_1 &: Jumper \\ event_1 &: High\_Jump \\ (event_1, moving\_object_1) &: hasPart \end{aligned}$$

- Why should we assume $object_3$ and $moving\_object_1$ are synonyms? And how?

Fusion as abduction

- If we added $object_3 = moving\_object_1$ the query

$$\{(X)_{[T_1,T_2]} \mid HJ\_Occurs_{[T_1,T_2]}(X), hasPart(X,Y), Famous(Y)\}$$

  would return a result.
- Which?
- Fusion as abduction: Add an assertion (same-as) such that the result set of a retrieval query is not empty

Also: Intra-modality fusion

- Domain knowledge: Good_Car(X) ← Car(X) ^ Works_well(X)
- Example (two sentences):
  - "This car has always been in a garage."
    - car-1 : Car, garage-1 : Garage
  - "It always worked well."
    - it-1 : Works_well
- Disambiguation required:
  - It is the car that worked well, not the garage
  - The car is indeed a good one (not a one that always was in a repair workshop)
- Fusion:
  - Gives rise to abducing car_1 = it_1
  - Such that { (X) | Good_Car(X) } has a result

Fusion as query computation

- If we had a query, abduction of same-as assertions could be used to formalize fusion
- What are the right queries?
- Where do they come from?
  - Look at concept definitions in the Tbox
  - Consider previously unanswered user queries

Queries again

- Initially: Pull technology
  - Queries for retrieving media objects
- Now: Queries for interpretation
  - Why not store the queries used for deriving interpretations as part of the metadata?
- Data arise in various contexts (as Aboxes)
- Use *context data for finding media objects*
  - “Matches” are all documents whose metadata contain queries that match a set of data (Abox assertions)
- This gives us: Push technology!
  - Based on context data, index relevant media objects (attached with queries that “match” on the data)
  - No need for the user to specify queries

Outlook

- Learn new interpretation rules in combination with corresponding Tbox axioms
- Deal with uncertainty in a systematic way
- Interaction with media objects in mobile interfaces

Interaction with media