Frame Detection over the Semantic Web

Bonaventura Coppola[1], Aldo Gangemi[2], Alfio Gliozzo[2], Davide Picca[3], Valentina Presutti[2]

coppola@disi.unitn.it, Università di Trento

{aldo.gangemi,alfio.gliozzo,valentina.presutti}@istc.cnr.it, ISTC-CNR, Rome

picca@unil.ch, Université de Lausanne

Outline

What

Why

How

What we do (1)

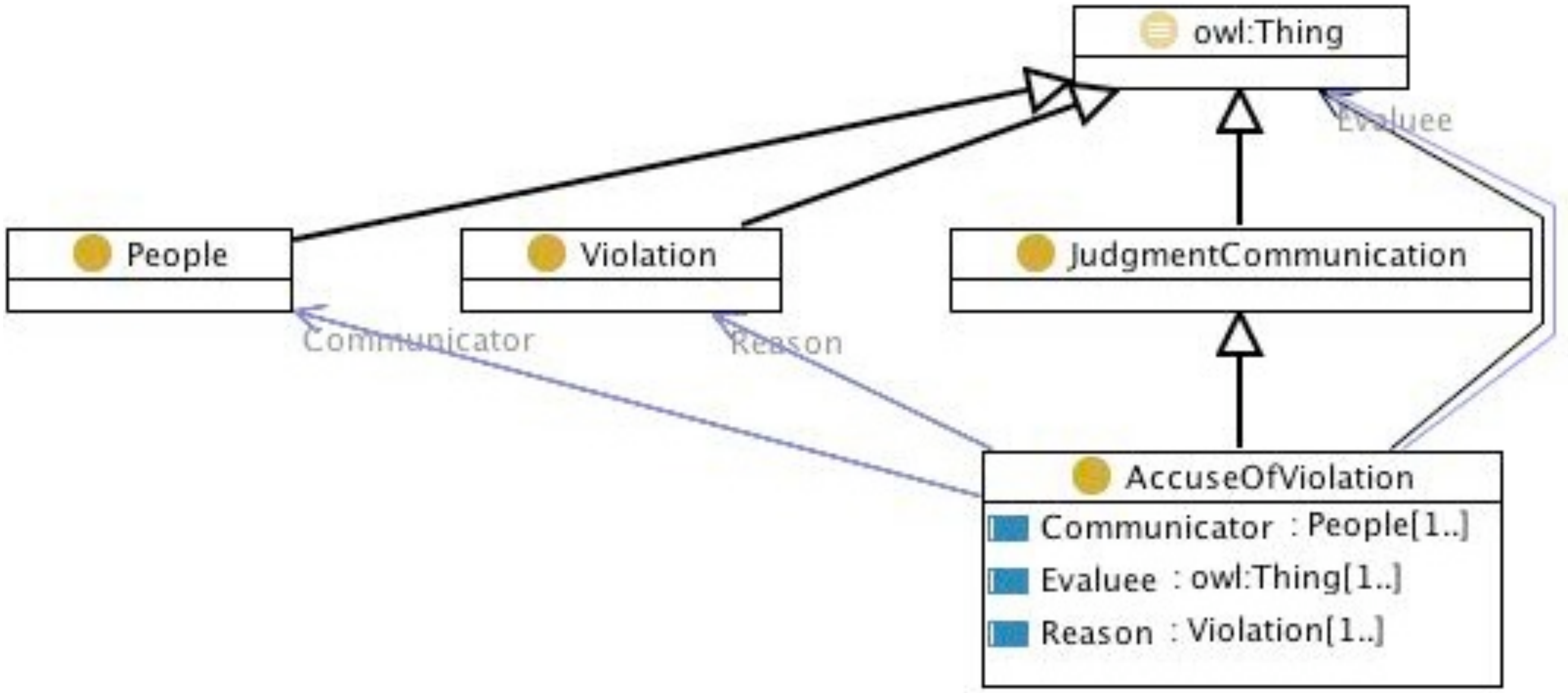
- This is an ontology learning experiment that reuses lexical frames from FrameNet, a textual corpus, and some OWL meta-models
- We learn OWL constructs with multiple related elements in the context of a frame, as occurring in a particular situation expressed in a text
- We use components
  - to identify occurrences of frames as expressed in text
  - to recognize the corresponding frame elements and their types
  - to learn domain-specific specializations and instantiations
  - to transform all that into ABox and TBox OWL ontologies

What we do (2)

- (S383) Public opinion and the press nowadays accuse us of being unavailable to give a response .
- FRAME: JudgmentCommunication
  - Communicator: "Public opinion" noun.cognition-Concept "and the press" noun.group-Concept"
  - Target: "accuse" verb.communication
  - Evaluee: "us" noun.group-Concept
  - Reason: "of being" verb.stative "unavailable" "to give" verb.possession "a response" noun.phenomenon-Concept"
- We learn domain frames (in this case: *public opinion accuses a group for unavailability*) by matching the frame element names to the corresponding super-senses, and by filtering out low-ranked results
- We transform domain frames into small reusable ontologies

STLAB
SEMANTIC TECHNOLOGY LABORATORY

A sample content pattern from a domain frame



Ontology learning: what structures?

- State of art: what do we learn?
  - instances (NER)
  - classes (NER, matching to lexica)
  - relations between instances
  - axioms: mostly taxonomic or disjointness, some work also on learning restrictions
  - entire ontologies, but as a collection of the previous ones

Expertise is more than that

- Evidence that units of expertise are larger than what we get with ontology learning
  – "Blinking" effects in reacting to events, in evaluating the actions and theories of the others, in understanding context, in interpreting news and ads, etc.
  – Competency questions
  – Task-based ontology evaluation
- We need to learn larger structures
  – but which ones, actually?

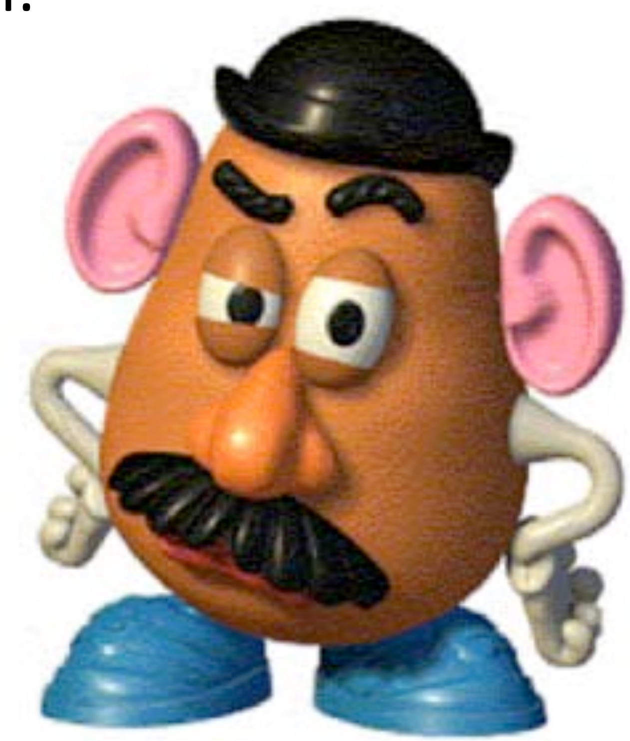
Does it really understand?

© guy & rodd, Dist. by UFS

The **driving** Frame

Driver: This is the being, typically human having a driving license, that controls the Vehicle as it moves.

Vehicle: This is the means of conveyance controlled by the Driver.



http://framenet.icsi.berkeley.edu/

Frame Detection

This frame deals with a Healer treating and curing an Affliction (the injuries, disease, or pain) of the Patient, sometimes also mentioning the use of a particular Treatment or Medication.

Medication [Med] **Semantic Type** Physical_entity	The injested, applied, injected, etc. substance designed to cure the Patient. He needs prolonged TREATMENT with anibiotics. Note the tight relationship between Treatment and Medicine.
Healer [Hlr] **Semantic Type** Sentient	The Healer, anyone who treats or cures the Patient, occurs as the External Argument of verbs: Doctors ALLEVIATED his suffering.

Cure

Healer

Patient

Medication

http://framenet.icsi.berkeley.edu/

STLAB
Semantic Technology Laboratory

Large structures in resources

- In resources
  - Sentences
  - Co-referential text fragments
  - Sub-categorization frames
  - Lexico-syntactic patterns
  - Lexical frames
  - Question patterns
  - Not only NL
    - Reengineering from Microformats, Infoboxes, Data models, Query types, etc.

Large structures in ontologies

- In ontologies
  - N-ary relations between instances
  - Classes with proper (and locally complete?) sets of restrictions
  - Content ontology design patterns

STLAB
SEMANTIC TECHNOLOGY LABORATORY

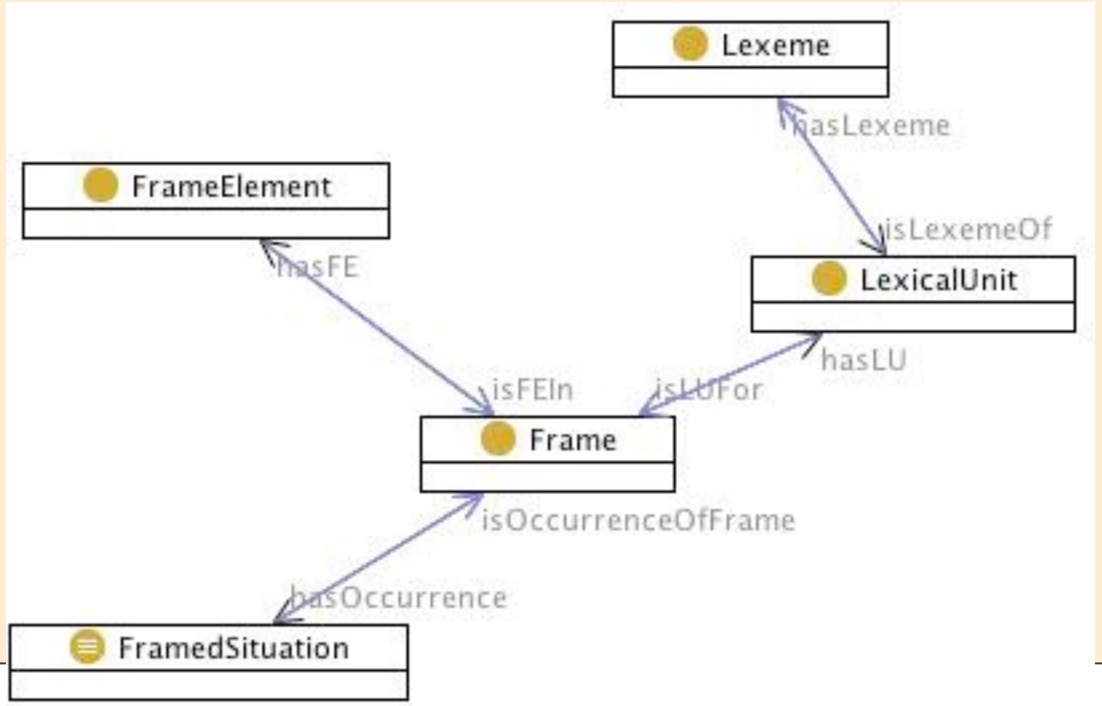
Frame Semantics and FrameNet

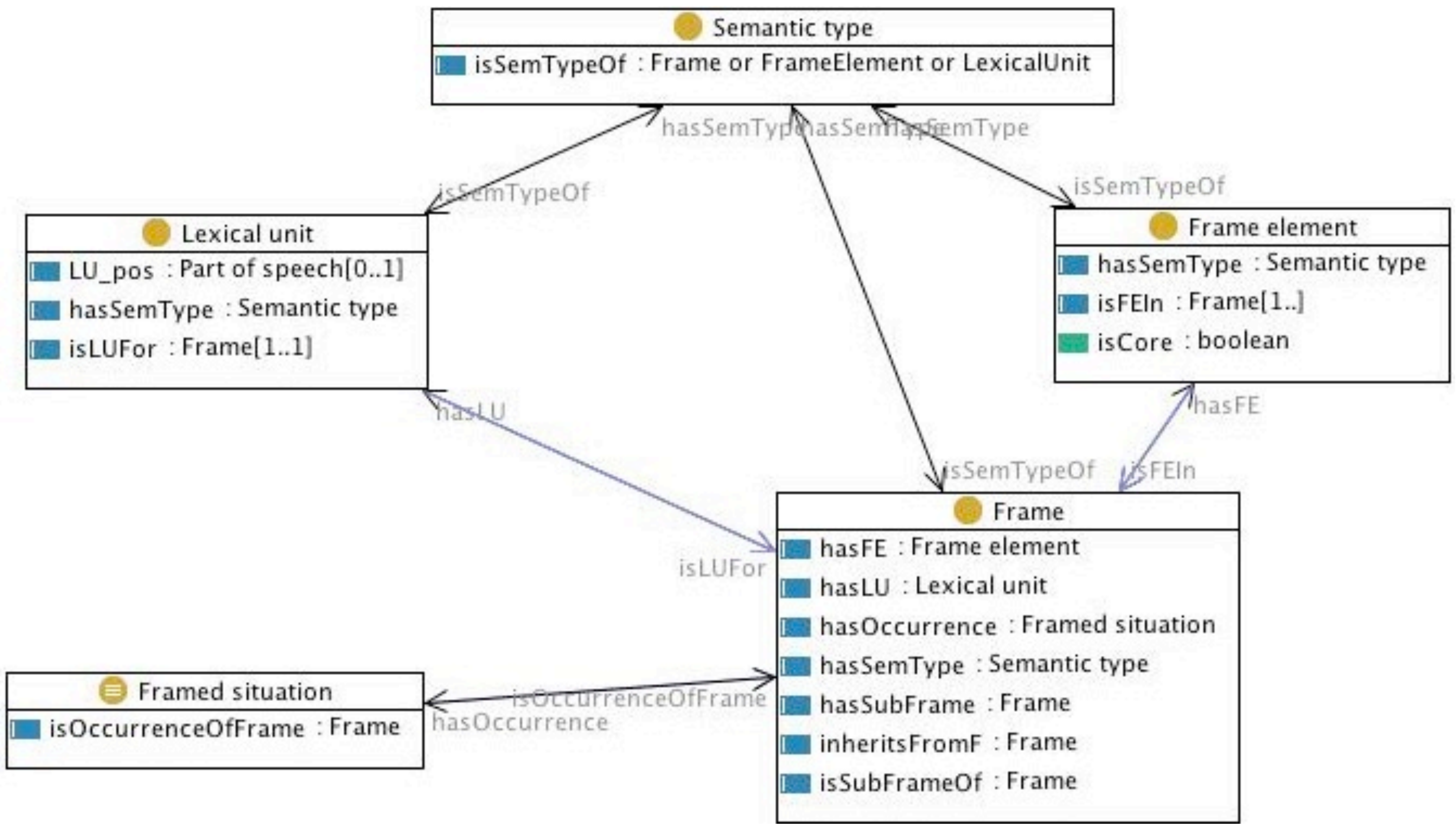
- (FrameNet) Frame: abstract representation of an event, situation, or property
- Example: COMMERCE_SCENARIO
  Core Frame Elements: BUYER, GOODS, MONEY, SELLER
  non-Core Elements: MANNER, MEANS, PURPOSE, RATE
  Lexical Units: buy, sell, purchase, acquisition, …
  Subframes: COMMERCIAL_TRANSACTION
- Instance: "Ralemberg said [he]SELLER already had a [buyer]BUYER [for the wine]GOODS"
- The FrameNet Project (University of Berkeley) developed definitions for ~800 Frames and ~4000 Frame-dependent Elements
- The FrameNet Data Base includes over 135,000 annotated sentences.
- FrameNet doesn't provide deep syntactic analysis (just shallow). Nonetheless, we currently heavily depend on deep syntax.

STLAB
SEMANTIC TECHNOLOGY LABORATORY

Our components and OntoFrameNet

- We have integrated two advanced text processing components: a *frame detector* and a *super-sense tagger*
- Segments of text including lexemes "evoking" lexical units and frame elements are collected for each frame
- They describe particular framed situations expressed in a text




Semantic type
isSemTypeOf : Frame or FrameElement or LexicalUnit
Lexical unit
LU_pos : Part of speech[0..1]
hasSemType : Semantic type
isLUFor : Frame[1..1]
Frame element
hasSemType : Semantic type
isFEIn : Frame[1..]
isCore : boolean
Frame
hasFE : Frame element
hasLU : Lexical unit
hasOccurrence : Framed situation
hasSemType : Semantic type
hasSubFrame : Frame
inheritsFromF : Frame
isSubFrameOf : Frame
Framed situation
isOccurrenceOfFrame : Frame
isSemTypeOf
hasSemType
hasLU
isLUFor
hasFE
isFEIn
isOccurrenceOfFrame
hasOccurrence

Frame detection (1)

- Method: Support Vector Machines (SVMs) as a general statistical machine learning framework, in which we plug in Tree Kernels to obtain a similarity measure capable of working over syntactic trees.
  - Semantic Role Labeling setting (Moschitti et al., CL, 2008)
  - Polynomial Kernel over handcrafted linguistic features (Palmer, Pradhan)
- Handcrafted features are very precise in capturing linguistic phenomena
- Tree Kernels are very practical for domain/language portability: no need of specific feature (re)engineering
- Best results are always achieved when the two approaches are exploited in combination

STLAB
SEMANTIC TECHNOLOGY LABORATORY

Frame detection (2)

- Steps:
  - (1) *Plain Text Preprocessing*, where Tokenization, POS-Tagging and deep constituency-based syntactic parsing are executed
  - (2) *Target Word Detection*, where the semantically relevant words bringing predicative information are detected (keywords or main verbs)
  - (3) *Frame Disambiguation*, where the correct frame for any target word is selected among several candidates
  - (4) *Boundary Detection*, where the sequences of words constituting the Frame Elements (arguments) are detected
    - binary classification problem over syntactic tree constituents
  - (5) *Role Classification*, which assigns semantic labels to the Frame Elements detected in the previous stage
    - frame-dependent multi-class classifier

STLAB
SEMANTIC TECHNOLOGY LABORATORY

Automatic Frame-based Annotation: example

- Plain text sentence (syntax omitted):

  Ralemberg said he already had a buyer for the wine.
- Target Word Selection (dictionary keyword: buyer)

  Ralemberg said he already had a <u>buyer</u> for the wine.
- Frame Disambiguation:

  Selected Frame: Commerce_Scenario
- Argument Boundary Detection:

  Ralemberg said [he] already had a [<u>buyer</u>] [for the wine].
- Argument Role Classification:

  Ralemberg said [he]SELLER already had a [<u>buyer</u>]BUYER [for the wine]GOODS.

Supersense tagging

- Supersenses are lexicographer's categories providing an initial broad classification for the lexicon entries in WordNet

  Guns_and_Roses$_{\text{-noun.group_instance}}$ plays$_{\text{–verb.communication}}$ at$_0$ the$_0$ stadium$_{\text{-noun.location_concept}}$

- The Super-Sense Tagger (SST, Ciaramita and Altun 2006) implements a Hidden Markov Model, trained with a perceptron algorithm (Collins 2002)
- It has been expoited for a deeper semantic interpretation of texts finalized to the specialization of frames in a domain perspective
- With SST, we identify types of frame elements in frame occurrences

STLAB
SEMANTIC TECHNOLOGY LABORATORY

Distilling frames with frequency thresholds

- Syntactic constraints
  - Occurrences of frame elements should only be nominals or noun phrases, while target lexemes can be verbs, nouns or adjectives.
- Type constraints
  - We count the occurrences of all different super-senses associated with the matched arguments, and we filter out those having less that 10 occurrences
    - For the frame Killing the possible types of the element Victim are: *group concept*: 90, *person concept*: 79, *person instance*: 41
- Redundancy constraints
  - We distill only those terms belonging to the concept types having more than 3 occurrences
    - Occurrences of the frame element Victim belonging to the *person.concept* super-sense include notions like child, woman and civilian, while occurrences categorized by the *noun.state* super-sense, e.g. recognition, have a sensibly lower frequency, and are most likely to be errors

Learning Domain Frames: Example

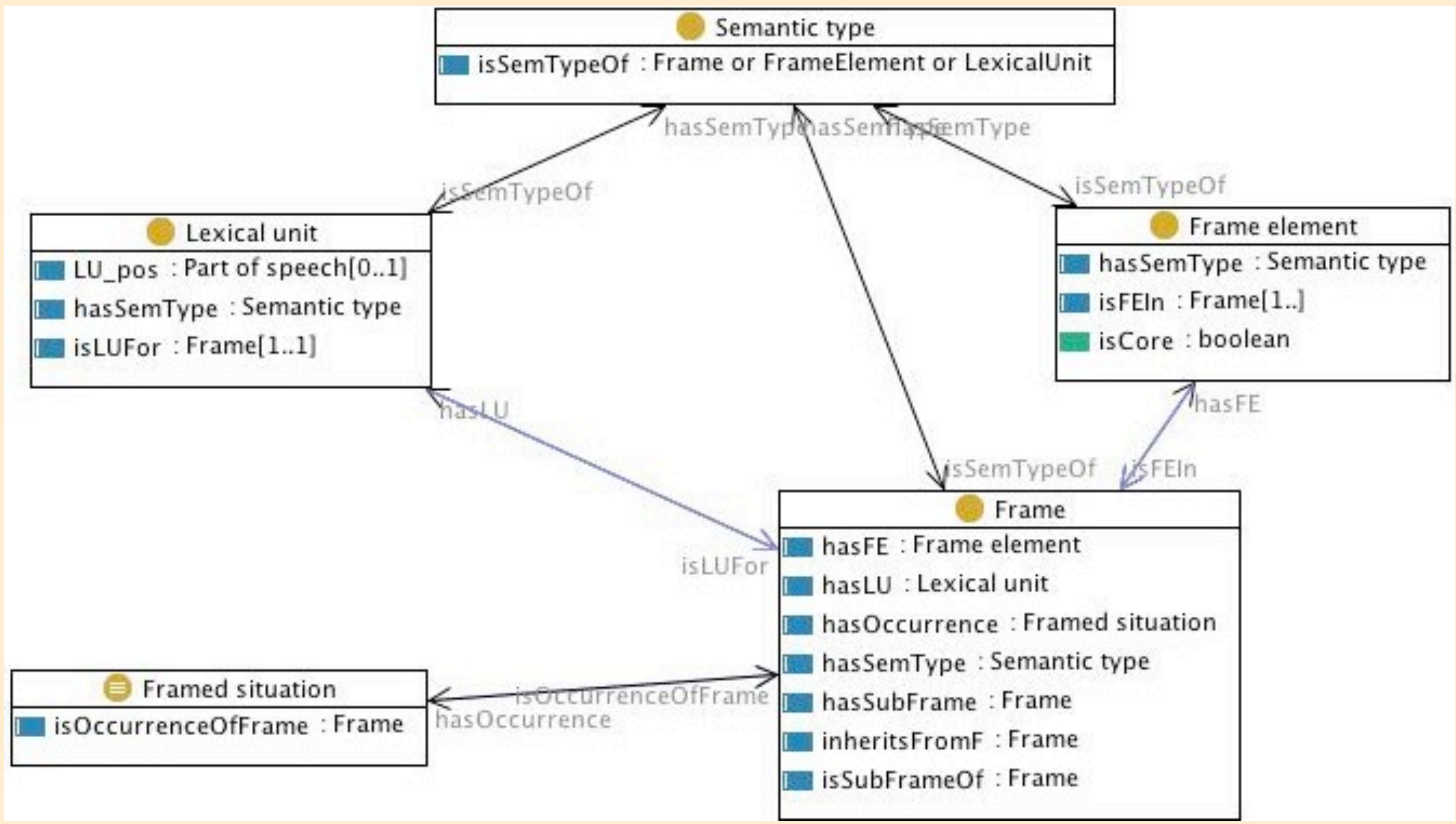
CommerceScenario

- Buyer:People_noun.group_Instance who read_verb.cognition this warning_noun.communication
- Target: buy_verb.possession
- Goods: these toys_noun.artifactInstance
- Recipient: for their children_noun.person_Concept or grandchildren_noun.person_Concept

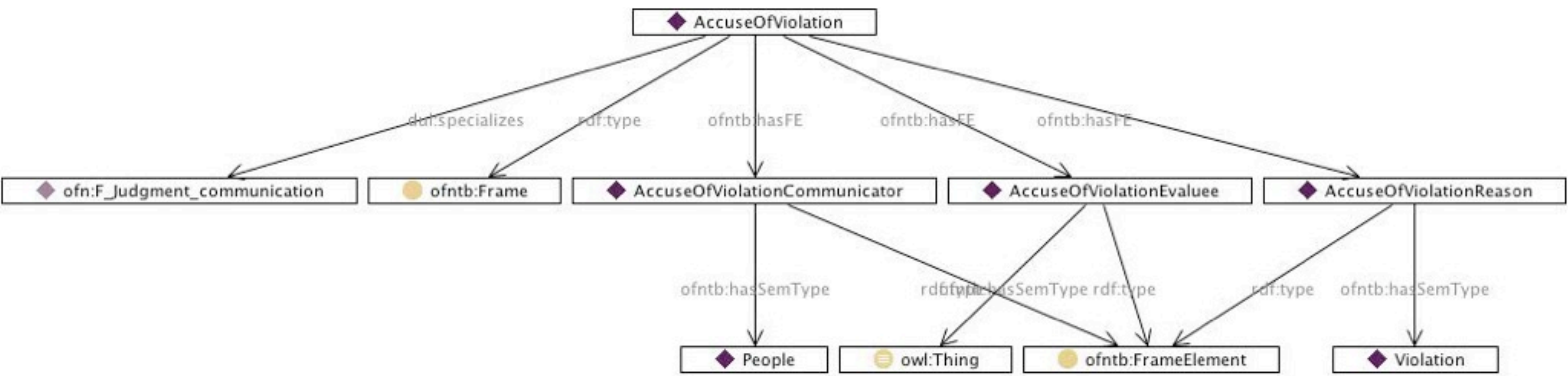
- **FRAME:** BuyingToysToChildren
  - **Buyer: [group]** People
  - **Goods [artifact]:** Toy
  - **Recipient [person]:** Child

Moving to OWL

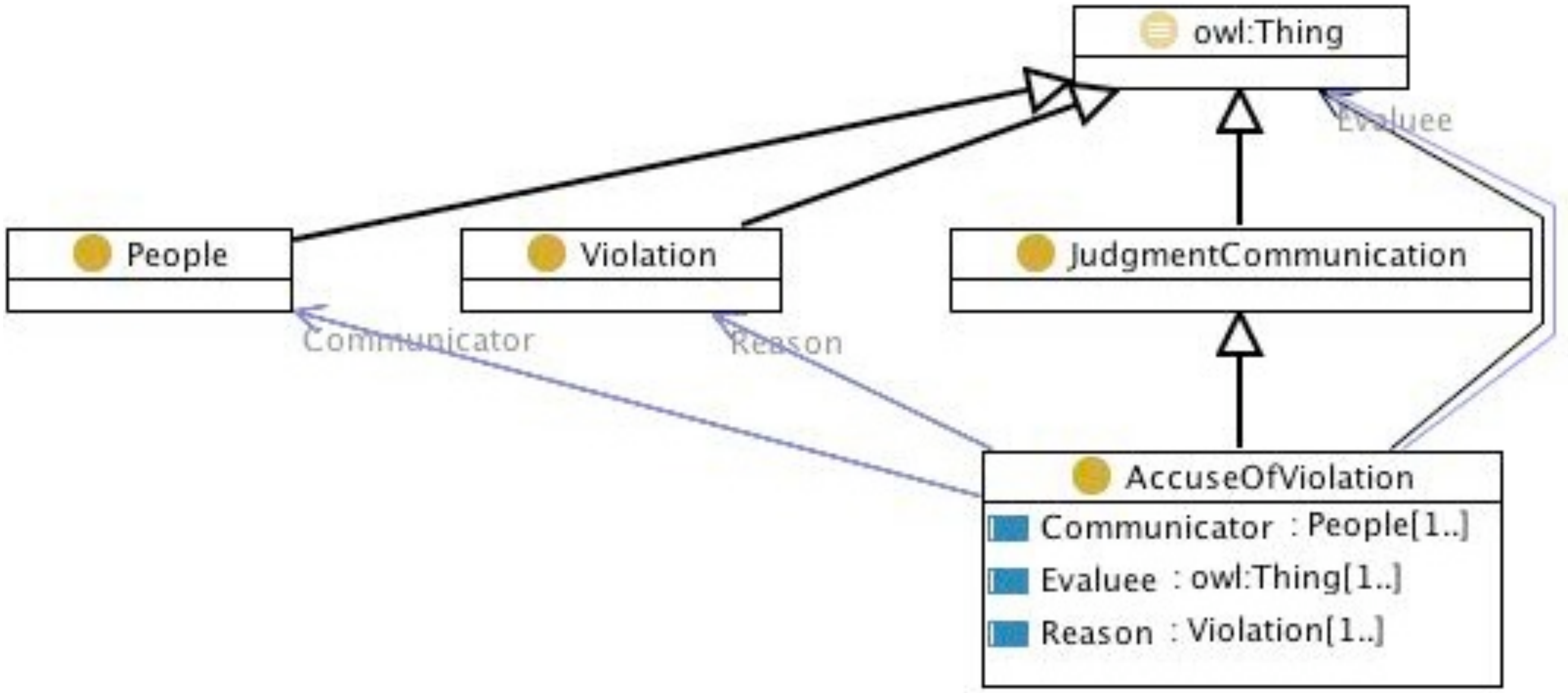
- How to represent a frame?
  - Bartlett, Fillmore, Davidson, Minsky, Brachman, Palmer
  - "Roles and their Descriptions", Masolo et al. KR2004, "What's in a Frame, Gangemi CUP2009
  - Content ontology design patterns, Presutti et al. 2008
  - Narayanan's ABox conversion of FrameNet
  - Scheffczyk's TBox conversion of FrameNet
  - OntoFrameNet, OntoVerbNet and LMM conversion (ABox+TBox reengineering strategy)


Semantic type
isSemTypeOf : Frame or FrameElement or LexicalUnit
Lexical unit
LU_pos : Part of speech[0..1]
hasSemType : Semantic type
isLUFor : Frame[1..1]
Frame element
hasSemType : Semantic type
isFEIn : Frame[1..]
isCore : boolean
Frame
hasFE : Frame element
hasLU : Lexical unit
hasOccurrence : Framed situation
hasSemType : Semantic type
hasSubFrame : Frame
inheritsFromF : Frame
isSubFrameOf : Frame
Framed situation
isOccurrenceOfFrame : Frame
isSemTypeOf
isSemTypeOf
isSemTypeOf
hasSemType
hasLU
isLUFor
hasFE
isFEIn
isOccurrenceOfFrame
hasOccurrence

A sample OntoFrameNet domain frame



A sample content pattern from a domain frame



Evaluation

- Frame Detector trained and evaluated on FrameNet corpus
  - Boundary Detection+Role Classification (words)
    - P=.747 R=.545 F1=.630
- SST trained and evaluated on SemCor corpus
    - P=.728 R=.683 F1=.699 (average through ss, stdev = ≈0.1)
- Frame-based ontology learning evaluated on Europarl corpus

STLAB
SEMANTIC TECHNOLOGY LABORATORY

The Europarl corpus

- Europarl corpus contains about 30M documents, extracted from the proceedings of the European Parliament.
- It includes 11 official languages of the European Union.
- We have used the English part, including ≈1.5M sentences and ≈40M words
- Sentences tagged with F, FE, LU, and SS

Results

- For three frames:
  - Killing, JudgmentCommunication, and Commerce
- ≈3k occurrences detected and tagged
- ≈300 occurrences distilled
- Transformation into LMM (OWL metamodel)
- Transformation into OWL
- Manual evaluation (only precision): P=.620

Related and Future work

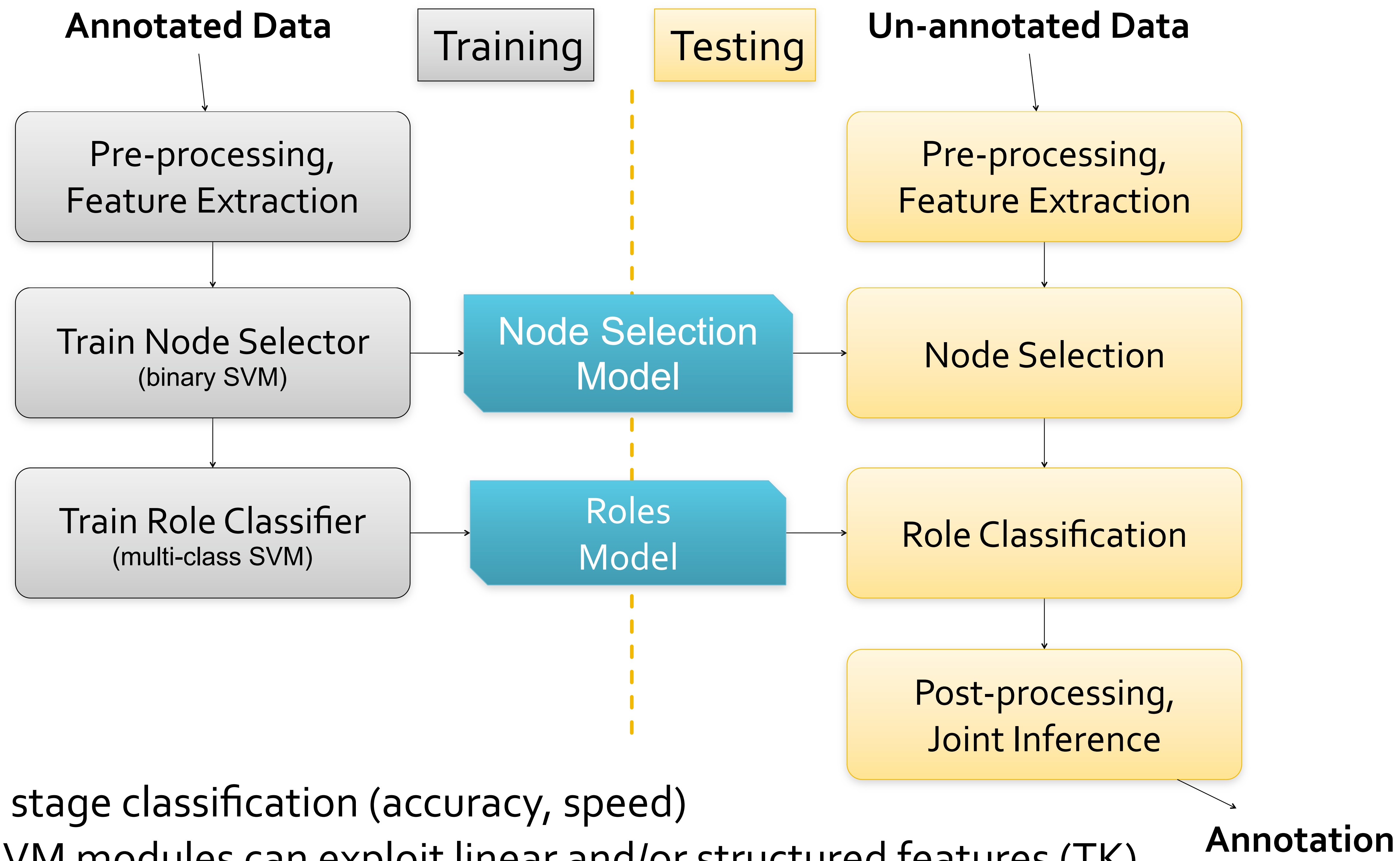
- Related frame learning work at University of Rome (Basili et al.)
- Critical aspects
  - low amount of "generic" lexical frames induce brittleness to very specialized frames
  - sparseness: texts do not encode expertise fully; domain frames are not filled completely with specialized concepts
- VerbNet, PropNet, etc. have similar limitations
- Foreseen extensions: mixing with ontology data, conceptual modelling patterns, community effort, non-linguistic resources

- Backup slides

System Architecture: Learning Model

- The FrameNet-based Annotation System exploits a Semantic Role Labeling setting (Moschitti et al., CL, 2008) based on:
  - Constituency-based syntactic annotation
  - Support Vector Machines
  - Polynomial Kernel over handcrafted linguistic features (Palmer, Pradhan: SRL state-of-the-art)
  - Several Tree Kernel variations (Subtree-, Subset-, Partial-TK)
- Handcrafted features are very precise in capturing linguistic phenomena
- Tree Kernels are very practical for domain/language portability (no need of specific feature (re)engineering)
- Best results are always achieved when the two approaches are exploited in combination

Machine Learning Setting for Traditional SRL



- 2 stage classification (accuracy, speed)
- SVM modules can exploit linear and/or structured features (TK)