Semantic Annotation in the Alvis Project

Adeline Nazarenko[2], Claire Nédellec[1],
Erick Alphonse[1], Sophie Aubin[2],
Thierry Hamon[2], Alain-Pierre Manine[1]

[1] Laboratoire Mathématique, Informatique et Génome (MIG), INRA

[2] Laboratoire d'Informatique de Paris-Nord (LIPN),
Université Paris-Nord & CNRS

Alvis project

Developing new technologies for distributed, topic-specific semantic-based search on internet

Query:

Author=*person:Crick* and *Author*=*person:Watson* and *Paper_title*=*title:The structure of DNA* and *Publication_date*=*date:1953*

Search for documents that comment the *famous* paper.

Answer : BBC news in 1953

"*in an article published **today** in Nature magazine, James D. **Watson** and Francis **Crick** describe the structure of a chemical called **d**esoxyribo**n**ucleic **a**cid,[..].*

Limitation of the keyword-based search

- Queries and search based on **keywords cooccurrences** do not exploit **semantic roles** (semantic types and relations).
- Although the simple cooccurrence of the four terms (*Crick*, *Watson*, *DNA structure*, *1953*) can be just spurious.

- Variations are not identified (*desoxyribonucleic acid* = *DNA structure* = *structure of DNA*)

- Individual terms may be semantically ambiguous (*Watson*).

Our framework

- Semantic search in Alvis relies on the **semantic annotation** of fined-grain semantic units and relations in the documents and their indexing.

- In specific domains, non-ambiguous annotation can be achieved by **linguistic analysis** and **domain-dependent resources**.

- Specific resources can be automatically acquired by **corpus-based machine learning methods**.

Annotation of semantic unit and relation requires linguistic processing

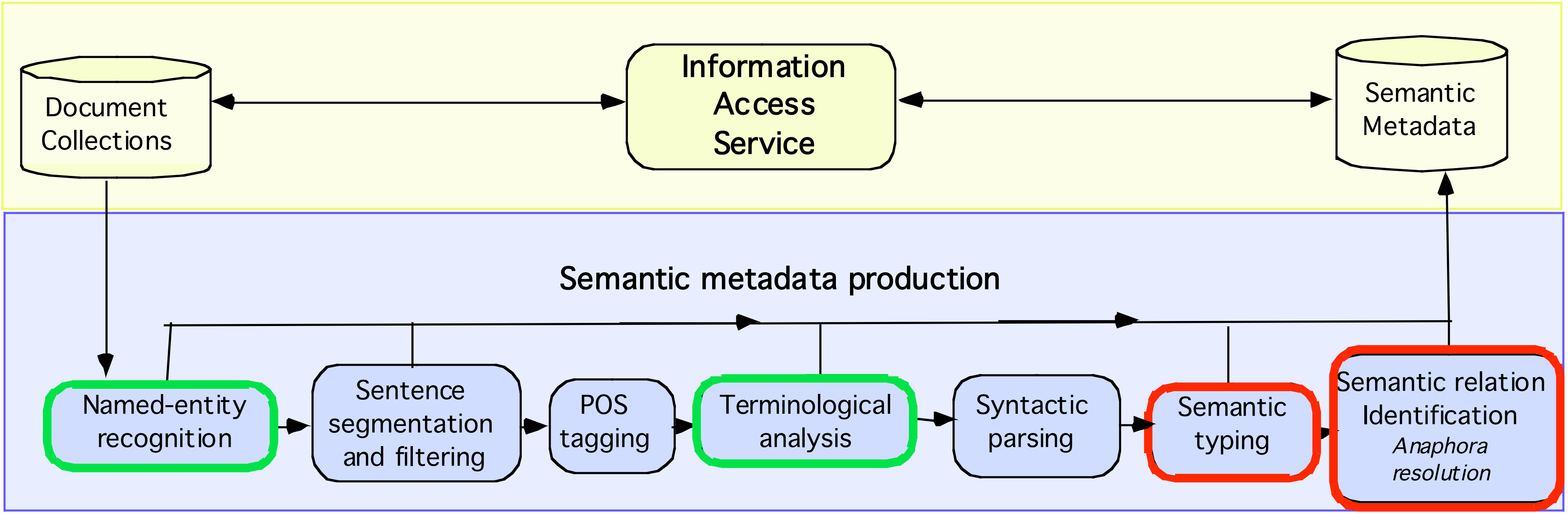
- The **semantic units** refer to the concepts and objects of the domain.
  - They do not always appear in their canonical form (variation and synonymy issues)
    Sigma K / sigma(K)
    Serum response element / Serum response factor
  - They may be ambiguous (polysemy issue)
    Has (both a gene and a verb)

The linguistic analysis of the semantic unit **morphology** and **contexts** solve these problems.

- Cooccurrence says little about the **semantic relations**

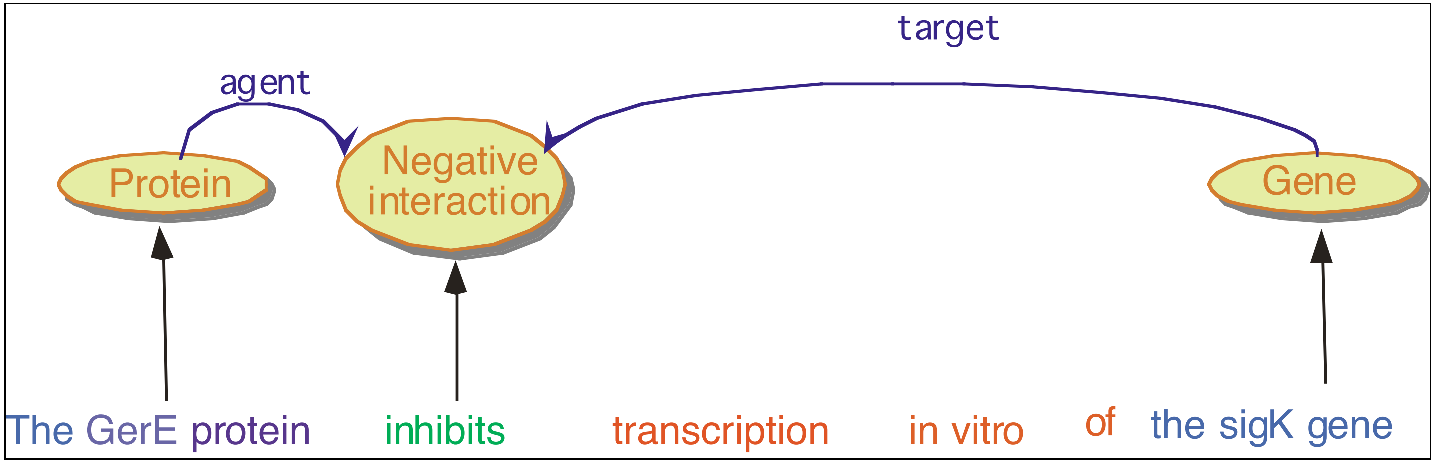
  GerE stimulates cotD transcription and cotA transcription […], and, unexpectedly, inhibits […] transcription of the gene (sigK) […]

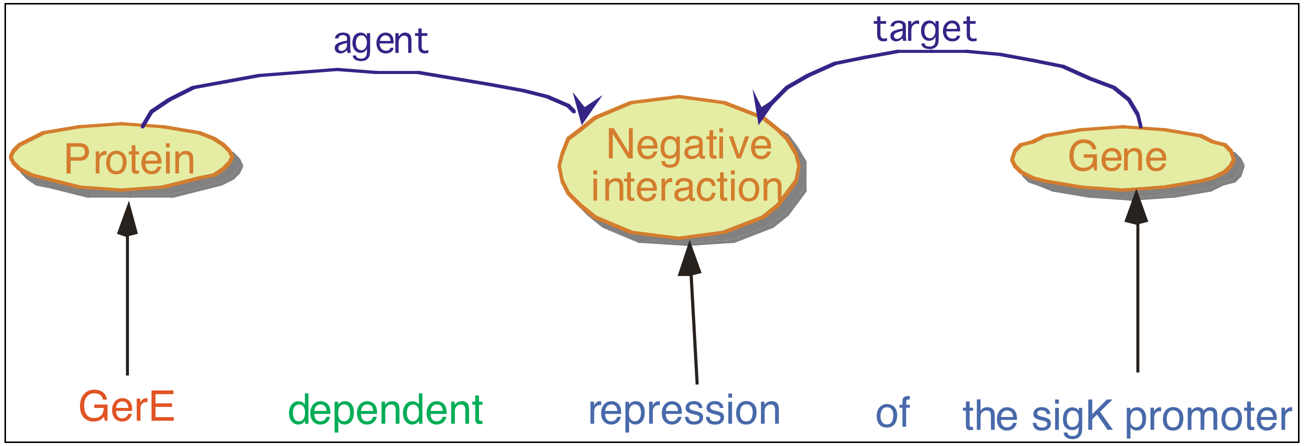
Semantic annotation with linguistic processing



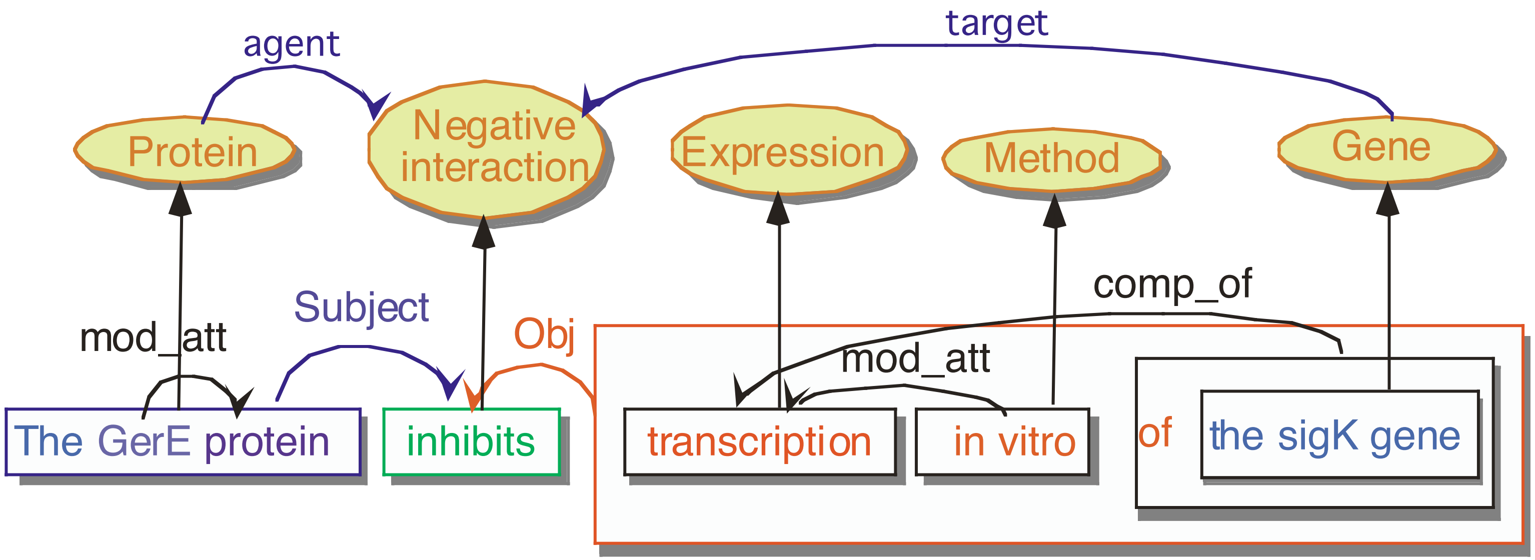
Semantic abstraction

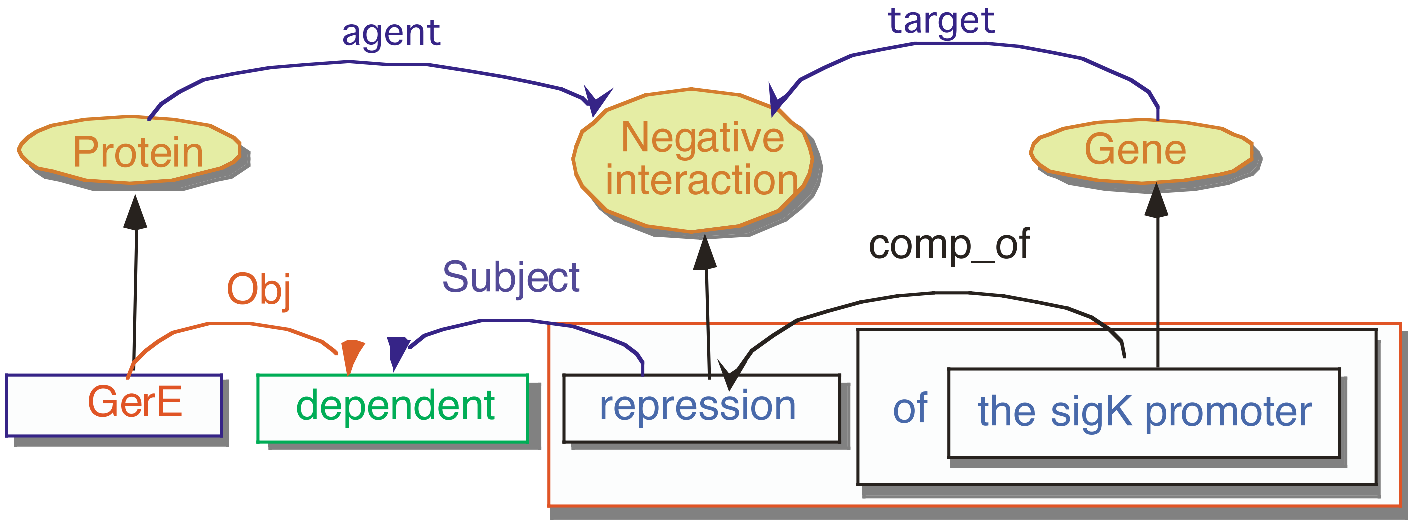
A same semantic representation of different formulations for efficient IR.



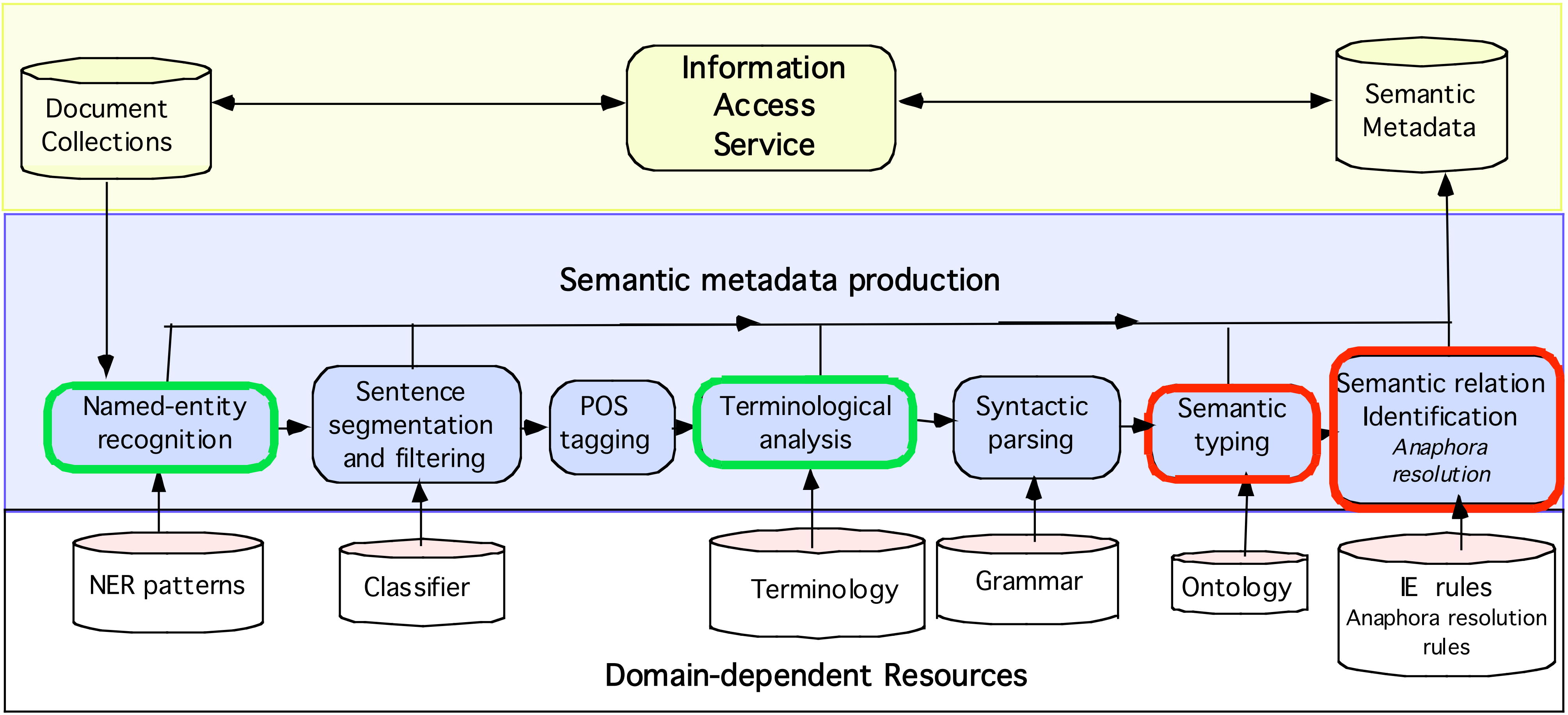


Linguistic analysis

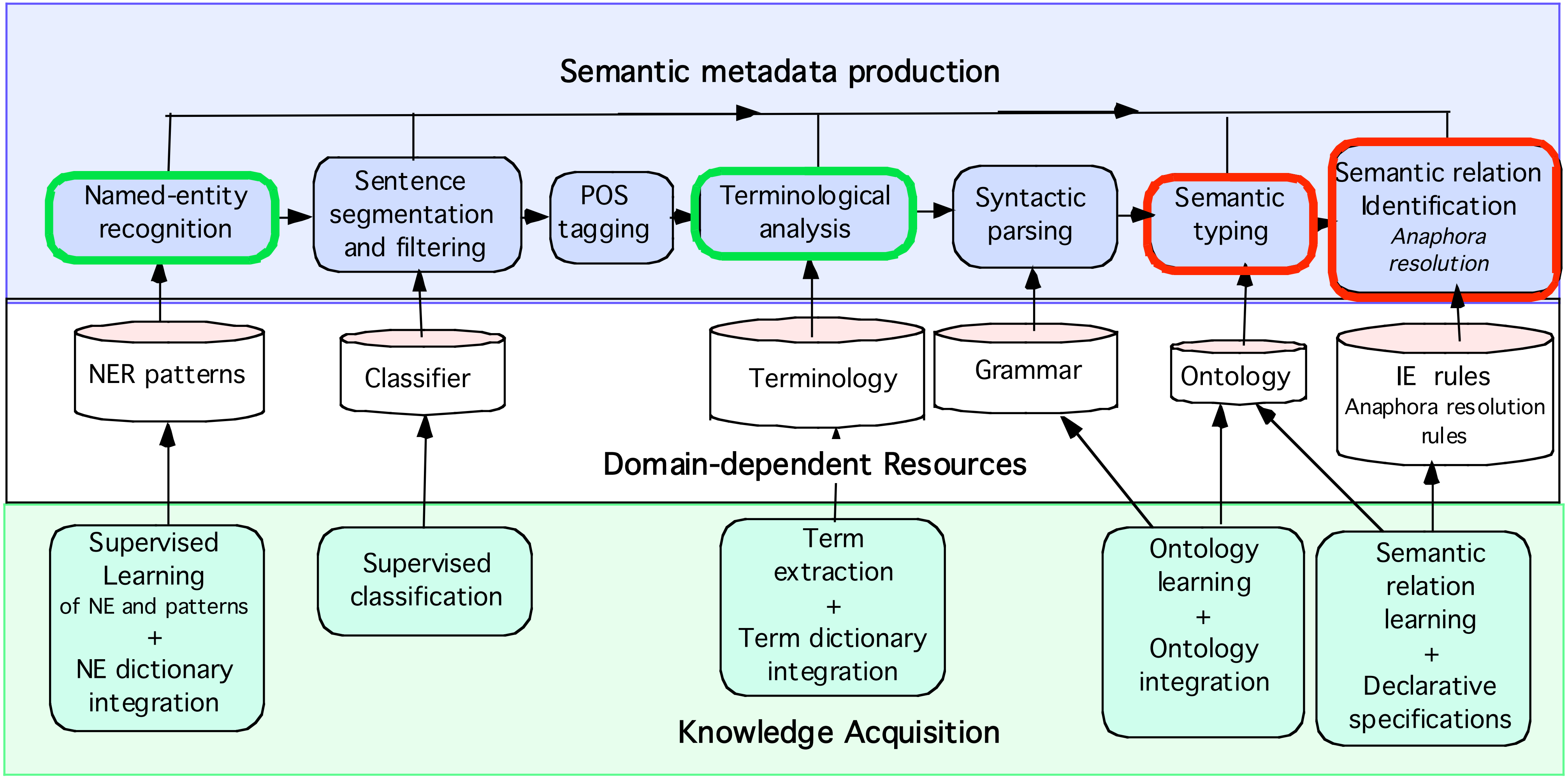




Specific resources are needed



Learning the resources


Semantic metadata production
Named-entity recognition
Sentence segmentation and filtering
POS tagging
Terminological analysis
Syntactic parsing
Semantic typing
Semantic relation Identification
Anaphora resolution
NER patterns
Classifier
Terminology
Grammar
Ontology
IE rules
Anaphora resolution rules
Domain-dependent Resources
Supervised Learning of NE and patterns + NE dictionary integration
Supervised classification
Term extraction + Term dictionary integration
Ontology learning + Ontology integration
Semantic relation learning + Declarative specifications
Knowledge Acquisition

Named-entity learning

Supervised learning for learning NER patterns of gene/protein names

In eight isolates of M. fermentans examined, malp occurred upstream of an operon encoding the phase-variable P78 ABC transporter;

Examples represented by linguistic features (mainly typographic).

- **First_upper**: the example is capitalized (^[A-Z])
- **Middle_upper**: the example contains a non-initial uppercase letter (^.+[A-Z])
- **Only_upper**: all letters of the example are uppercase? (^[A-Z]*$)
- **Last_digit**: the last character of the example is a digit? ([0-9]$)

...

Experimental results

	Precision	Recall
C4.5	92,5	91,6
NB	88,6	73,4

Best **NLPBA**: Precision 76% Recall 69,4%

BioCreative: 83% Recall-Precision

Terminology acquisition by YaTea

YaTea term acquisition tool combines *existing terminology* matching (good precision) and *corpus-based term extraction* (good coverage).

Input

Training corpus tagged with POS information and existing terminology

During[ADV] sporulation[NOUN] of[PREP] Bacillus subtilis[P-NOUN], spore[NOUN]

Method

1.Corpus chunking based on frontier category detection

During / sporulation of Bacillus subtilis /, / spore coat proteins / encoded by /

2. Recursive parsing of chunks according to
   - Syntactic patterns `NOUN NOUN`
   - Forbidden structures and subcomponents (*of course*)
   - Specific patterns of certified terms (*in vitro)*
   - Generation of term variants using morpho-syntactic rules

     `NOUN1 NOUN2 = NOUN2 of NOUN1`

Examples of term tagging

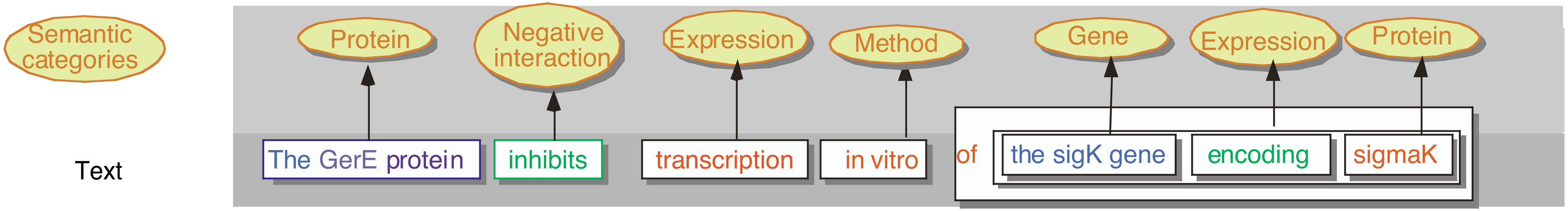
Existing terminology: Gene Ontology terminology mapping (in green)

/Combined/ /action/ /of/ /two/ /transcription/ /factors/ /regulates/ /genes/ /encoding/ /spore/ /coat/ /proteins/ /of/ /Bacillus/ /subtilis/ /.//During/ /sporulation/ /of/ /Bacillus/ /subtilis/ /,/ /spore/ /coat/ /proteins/ /encoded/ /by/ /cot/ /genes/ /are/ /expressed/ /in/ /the/ /mother/ /cell/ /and/ /deposited/ /on/ /the/ /forespore/ /.//transcription/ /of/ /the/ /cotB/ /,/ /cotC/ /,/ /and/ /cotX/ /genes/ /by/ /final/ /sigma/ /(/ /K/ /)/ /RNA/ /polymerase/ /is/ /activated/ /by/ /a/ /small/ /,/ /DNA/-/binding/ /protein/ /called/ /GerE/ /.//The/ /promoter/ /region/ /of/ /each/ /of/ /these/ /genes/ /has/ /two/ /GerE/ /binding/ /sites/ /./

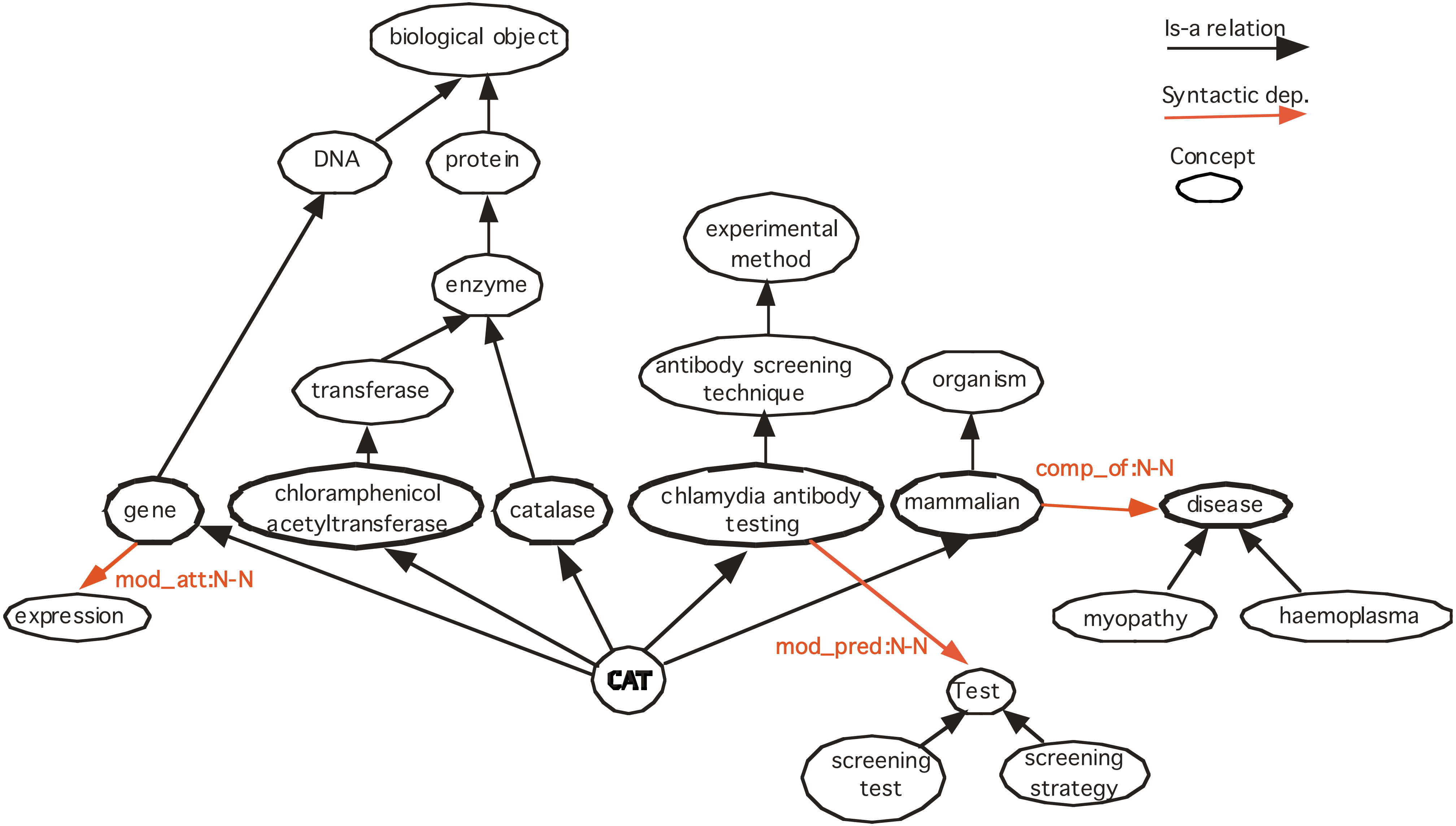
YaTea term mapping (in green)

/Combined/ /action/ /of/ /two/ /transcription/ /factors/ /regulates/ /genes/ /encoding/ /spore/ /coat/ /proteins/ /of/ /Bacillus/ /subtilis/ /.//During/ /sporulation/ /of/ /Bacillus/ /subtilis/ /,/ /spore/ /coat/ /proteins/ /encoded/ /by/ /cot/ /genes/ /are/ /expressed/ /in/ /the/ /mother/ /cell/ /and/ /deposited/ /on/ /the/ /forespore/ /.//Transcription/ /of/ /the/ /cotB/ /,/ /cotC/ /,/ /and/ /cotX/ /genes/ /by/ /final/ /sigma/ /(/ /K/ /)/ /RNA/ /polymerase/ /is/ /activated/ /by/ /a/ /small/ /,/ /DNA/-/binding/ /protein/ /called/ /GerE/ /.//The/ /promoter/ /region/ /of/ /each/ /of/ /these/ /genes/ /has/ /two/ /GerE/ /binding/ /sites/ /./

Semantic type tagging



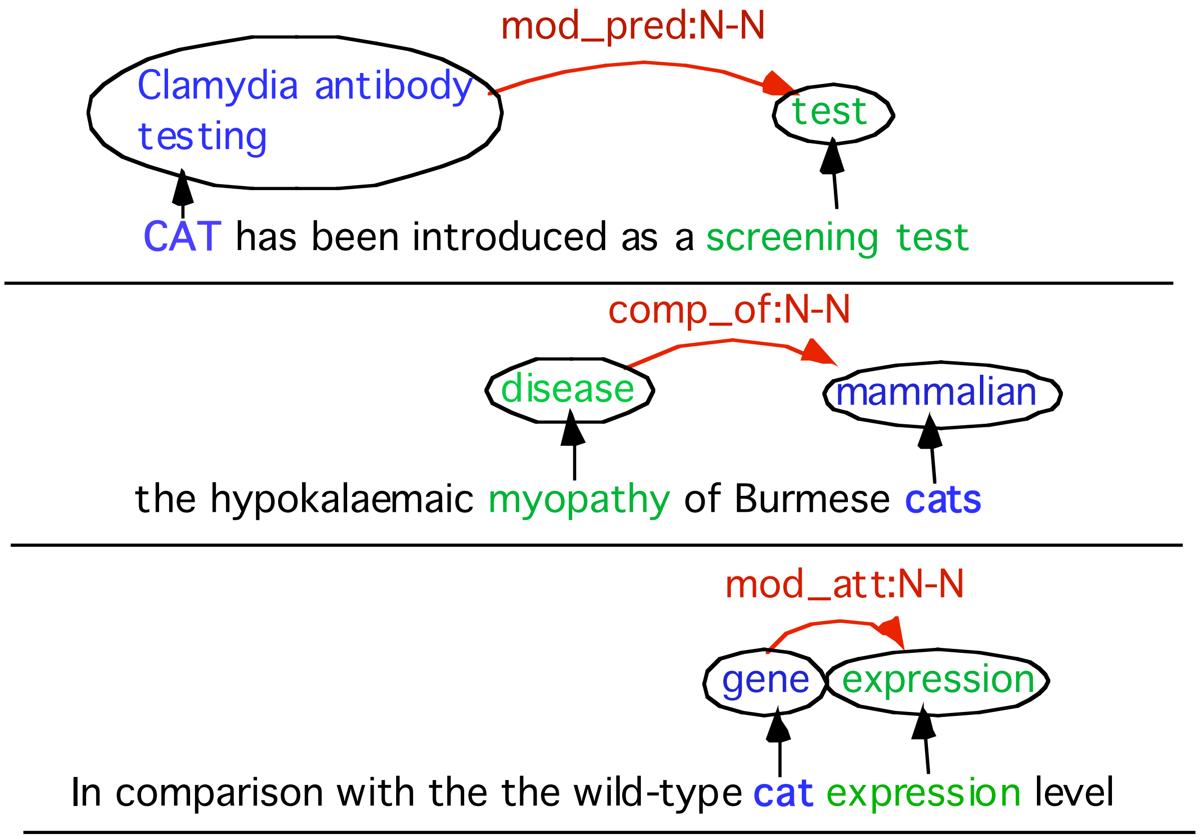
Semantic type learning by Asium



Semantic disambiguation with syntactic context

Given,

- *Restrictions of selection* associated to the concepts of the ontology
- *Is-A hierarchies*

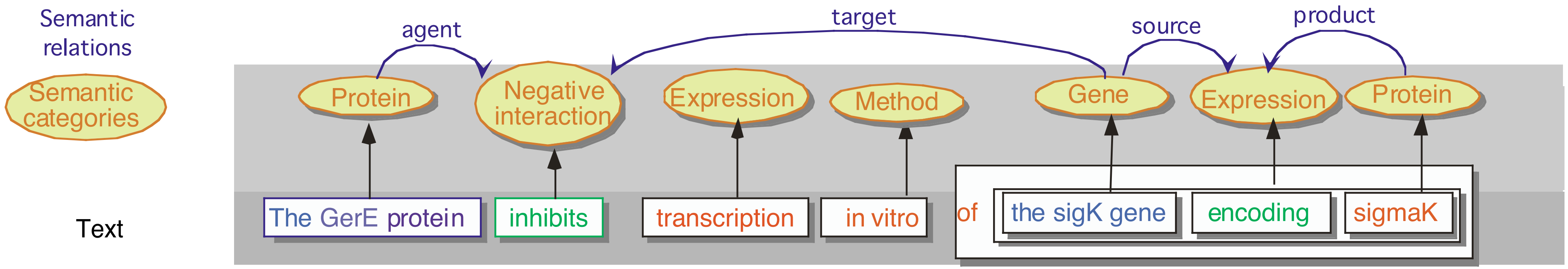


CAT has been introduced as a screening test

the hypokalaemaic myopathy of Burmese cats

In comparison with the the wild-type cat expression level

Tagging semantic relations



Rules for semantic relation annotation

GerE stimulates cotD transcription and cotA transcription [...], and, unexpectedly, inhibits [...] transcription of the gene (sigK) [...]

Example of information extraction rule

interaction (X,Z):-

is-a(X,protein), subject(X,Y), cat(Y,verb), is-a(Y,interaction), cat(Z,NP), obj(Z,Y), is-a(Z,gene-expression).

Interpretation

If the subject X of an interaction verb Y is a protein name, and the object Z is a gene expression,

then, X is the agent and Z is the target of the interaction

Rule learning with Propal (*ILP-based*)

Learning method

Supervised relational learning,
Horn clauses
Multi-class learning: top-down ILP method Propal [Alphonse, 2003]

Training data pre-processing

1. Selection of relevant documents.
2. Segmentation and filtering of relevant sentences.
3. Manual annotation of the relations in the positive training data.
4. Negative example generation (near-miss selection in relevant sentences under closed-word assumption)
5. Training example preprocessing (linguistic processing and saturation by BK).

Application of the learning method for acquiring the rules representing the discriminant linguistic attributes.

- "Learning Language in Logic" challenge (*ICML 05 LLL workshop*) see webpage.

Preliminary results on relation learning

- Training data: gene interactions (agent, target) in *Bacillus subtilis*
  LLL challenge dataset on "action without coreference"
- Linguistic normalization (lemma and syntactic relations) and abstraction
- Rule learning with **Propal**

	Recall	Precision	F-measure
[Goadrich et al., 2005], data without linguistics	80,6	42,6	58,5
[Riedel and Klein, 2005] data with linguistics	52,8	86,4	65,5
[Propal] linguistics + semantic abstraction	61,8	63,6	62,7

Conclusion

Semantic annotation of free text in specialized domains is a complex task with high added-value

2 complementary approaches

- **Shallow and statistics-based processing**

➔ Easy to design
➔ The information retrieved is partially noisy

- **Text normalization and Machine Learning**

➔ Saves time of adaptation of the resources to the task
➔ Better coverage of the diversity of the linguistic expressions
➔ Complex architecture, difficult to design