VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Unsupervised and semi-supervised learning of action ontology using domain-specific corpora

Daiva Vitkutė-Adžgauskienė, Irena Markiewicz, Tomas Krilavičius

Faculty of Informatics & Centre of Computational Linguistics
Vytautas Magnus University

Kaunas, 2013

The goals

The goals of this presentation are:

- To show the possibilities of **Natural Language Processing (NLP) in building action ontology** from domain-specific corpora
- To show how **unsupervised and supervised ontology learning methods can be combined** for action ontology building
- To **build a framework** for action ontology inducing from domain-specific texts
- To give **some examples** from experiments with a crawled Chemlab corpus

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Meta-model for ontology learning

General metodology for ontology building from texts can be described using the following model (Philipp Cimiano, 2006):

$$M=\{D, LA, T, S, C, TR\},$$

- where *D* is the text corpus documents collection,
- LA – linguistic annotations, i.e. metadata for corpus texts,
- *T* - terminology collection,
- *S* - synonym collection,
- *C* – ontology concepts collection,
- *TR* – ontology relations.

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Unsupervised vs supervised ontology learning

- **Ontology learning** – integration of different disciplines and tools in order to facilitate ontology construction.
- The main points of ontology learning framework:
  - **Input** (prior knowledge – texts, preprocessed texts, dictionaries, other ontologies, …)
  - **Learning methods** (unsupervised vs supervised, statistical vs logical, etc.)
- **Unsupervised ontology learning** – ontology concept and association extraction from (preprocessed) texts
- **Supervised ontology learning** – additional labeled information, structured semantical information and lexical patterns are used

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Unsupervised ontology learning

- Different NLP methods are applied to domain text collections – domain corpora:
  - term extraction,
  - collocation extraction,
  - named entity recognition,
  - word space model for similarity checks
- Corpora may be linguistically preprocessed – morphological annotations, dependency parses added
- Statistical (prevailing) and rule-based technologies are used

Supervised ontology learning

- Additional labeled information is used to facilitate ontology learning:
  - Semantically annotated training corpora
  - Structured linguistic databases (e.g. WordNet semantic taxonomy)
  - Lexical-syntactic search patterns (e.g. for hypernym/hyponym extraction: *NP0* such as *{NP1, NP2…,(and|or) }NPn*)

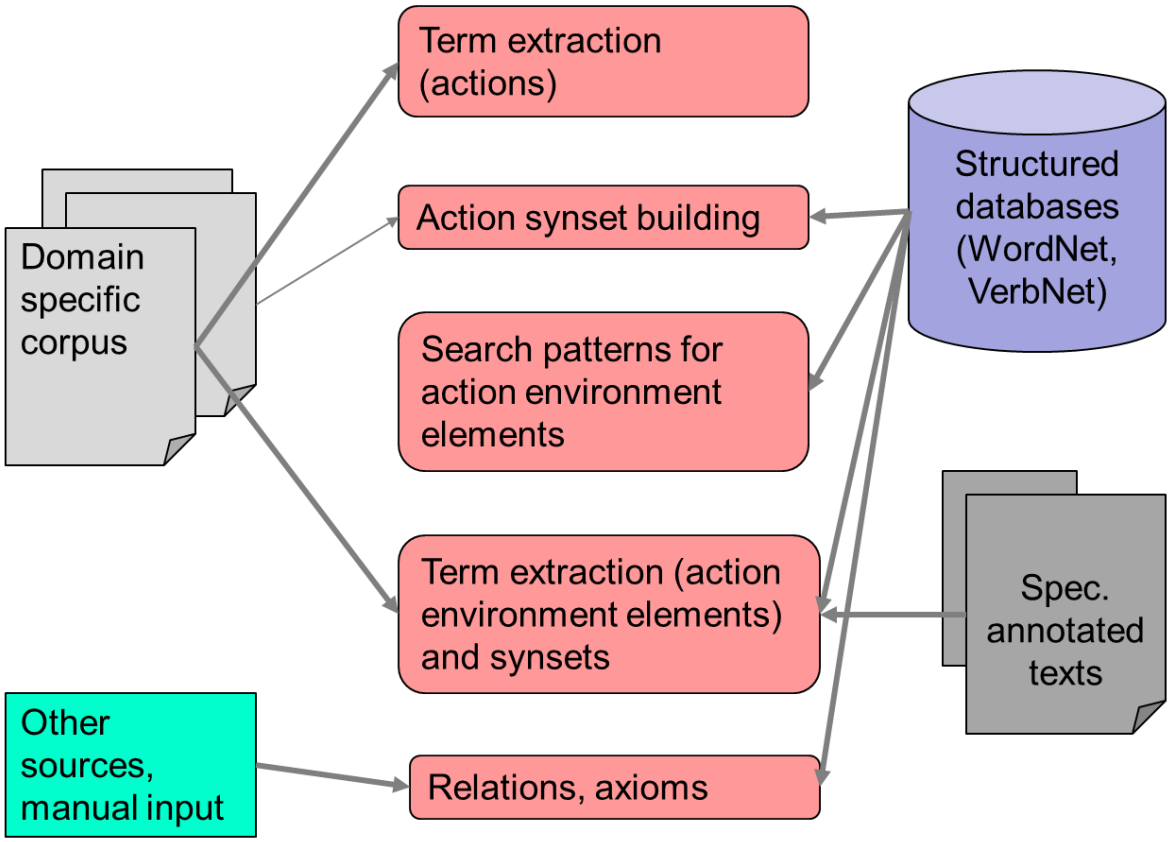
VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Unsupervised vs supervised ontology learning – pros and cons

- **Unsupervised** ontology learning methods:
  - are convenient for extraction processes **automation (+)**
  - require **large representative corpora** at the input **(-)**
- **Supervised** ontology learning methods:
  - works with **smaller text corpora (+)**
  - requires **manual effort** for producing labeled data and search patterns **(-)**
- **Combinations of both methods** are used for compromise between automation level and output quality

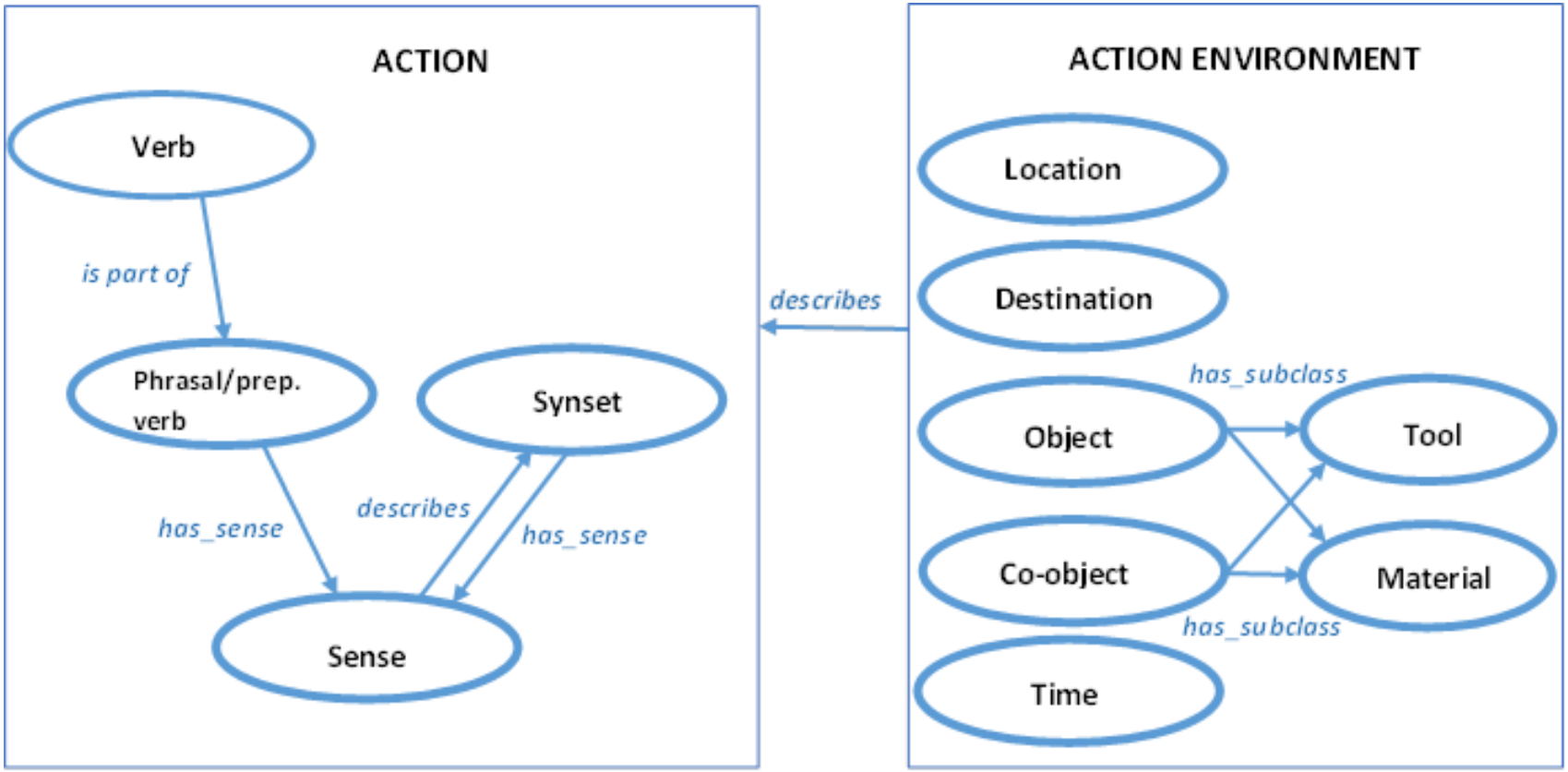
Combined approach for action ontology learning

- A combined approach for action ontology extraction encompasses the following steps:



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Action ontology – conceptual model



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Experimental data for illustration

- Input – domain corpus (collected from online texts on chemistry lab processes);
  - Collected data describes chemistry lab experiments, basic rules, instruments and techniques.
- Crawled texts were cleaned and filtered (corpus contains only texts, which are longer than 1600 symbols). Result – chemistry lab corpus of 1 971 415 words.
- Texts were morphological annotated and lemmatized

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

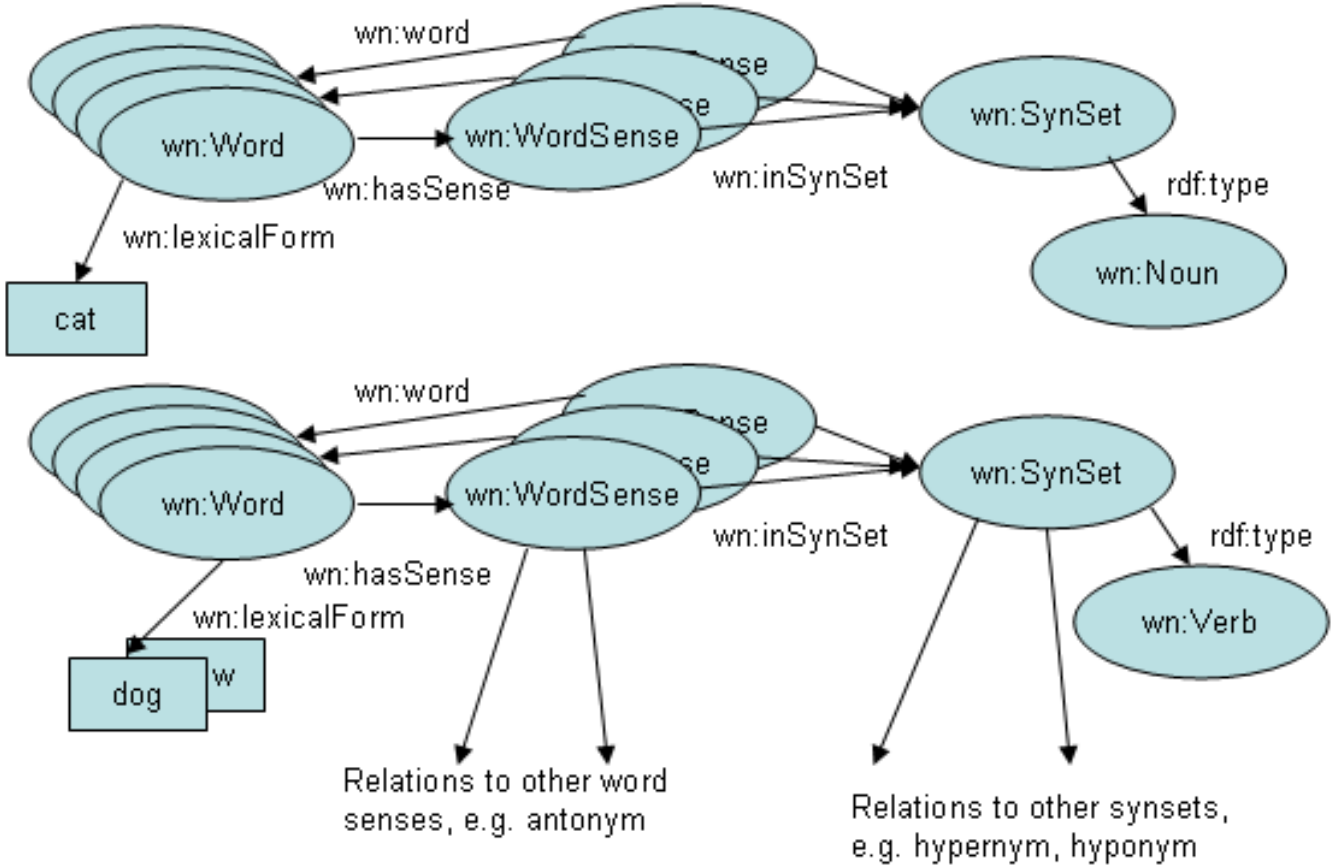
Supplemenary datasets - Wordnet

- Lexical database for English language
- Contains: nouns, verbs, adjectives, adverbs
- Describes relations:
  - for nouns (*hypernyms, hyponyms, holonyms, meronyms*);
  - for verbs (*hypernyms, troponyms, etc.*);
  - for adjectives (*relativeness, similarity, participation*);
  - for adverbs (*common adjectival core*).

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Supplemenary datasets - Wordnet

WordNet relations



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Supplemenary datasets - VerbNet

- VerbNet is a domain independent Verb lexicon for English language with extended conceptual classes (Beth Levin classification, based on *alternation* – the syntactic behavior of verbs)
- Includes mappings to other lexical resources: WordNet, Xtag and FrameNet
- Structure:
  - roles and restrictions: actor, agent, attribute, location, destination, source, instrument, material, product, patient, predicate, recipient, time;
  - members of verb semantic group;
  - frames with common examples and syntax structure.

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Supplemenary datasets - VerbNet

VerbNet thematic roles (part I)

Role	Description
Agent:	generally a human or an animate subject
Attribute:	attribute of Patient/Theme refers to a quality of something that is being changed
Destination:	end point of the motion, or direction towards which the motion is directed
Source:	start point of the motion
Location:	underspecified destination, source, or place, in general introduced by a locative or path prepositional phrase.
Actor:	Used for some communication classes: meet, marry

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Supplemenary datasets - Verbnet

VerbNet thematic roles (part II)

Role	Description
Product:	end result of transformation.
Patient:	used for participants that are undergoing a process or that have been affected in some way.
Recipient:	target of the transfer. Used by some classes of Verbs of Change of Possession, Verbs of Communication, and Verbs Involving the Body.
Time:	class-specific role, used in Begin-55.1 class to express time.
Instrument:	used for objects (or forces) that come in contact with an object and cause some change in them. Generally introduced by a `with' prepositional phrase.
Material:	Start point of transformation

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Supplemenary datasets - Verbnet

VerbNet data structure example

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)

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Terminology extraction (actions)

- Rule-based technologies
- Term-specific linguistic patterns:
  - Verbs
  - Propositional verbs (verb + proposition)
  - Phrasal verbs (verb + [direct object] + adverb)
- Input data: morphologically annotated corpus
- Term frequencies and input from reference information sources can be used for filtering rare terms

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Experiment – term extraction (actions)

Put (1143)

VB+IN	freq
put in	353
put on	149
put into	111
put of	94
put away	43
put back	32
put to	29
put off	16
put out	15
put at	12
put down	11
put as	8
put from	4

Mix (1640)

VB+IN	Freq
mix with	342
mix of	189
mix to	178
mix together	99
mix for	53
mix up	45
mix into	30
mix at	24
mix as	23
mix until	21
mix after	16
mix under	16
mix by	13
mix on	8
mix over	5

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Action senses

- Action title can be the same for several action senses.
- Example from WordNet (senses for “remove”):

Sense 1
remove, take, take away, withdraw -- (remove something concrete, as by lifting, pushing, or taking off, or remove something abstract; "remove a threat"; "remove a wrapper"; "Remove the dirty dishes from the table"; "take the gun from your pocket"; "This machine withdraws heat from the environment")

Sense 2
remove -- (remove from a position or an office)

Sense 3
get rid of, remove -- (dispose of; "Get rid of these old shoes!"; "The company got rid of all the dead wood")

Sense 4
take out, move out, remove -- (cause to leave; "The teacher took the children out of the classroom")

Sense 5
remove, transfer -- (shift the position or location of, as for business, legal, educational, or military purposes; "He removed his children to the countryside"; "Remove the troops to the forest surrounding the city"; "remove a case to another court")
    => transfer, shift -- (move around; "transfer the packet from his trouser pockets to a pocket in his jacket")

Sense 6
absent, remove -- (go away or leave; "He absented himself")
    => disappear, vanish, go away -- (get lost, as without warning or explanation; "He disappeared without a trace")

Sense 7
murder, slay, hit, dispatch, bump off, off, polish off, remove -- (kill intentionally and with premeditation; "The mafia boss ordered his enemies murdered")
    => kill -- (cause to die; put to death, usually intentionally or knowingly; "This man killed several people when he tried to rob a bank"; "The farmer killed a pig for the holidays")

Sense 8
remove, take away -- (get rid of something abstract; "The death of her mother removed the last obstacle to their marriage"; "God takes away your sins")

Action synsets

- **Synset** (a synonym ring) is a group of semantically equivalent data elements
- Synset information is well represented in WordNet lexical database (semantic taxonomy)
- **An action synset** can be built **by automated analysis of WordNet senses** for corresonding action verbs
  - Inadequate senses filtered out using Word Space Model (WSM)
  - Different verbs with the same sense grouped – **synset growing**
- Other external sources with reference information (e.g. VerbNet) can be used for synset growing

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Experiment – synset growing ("put")

Initial Action	Synset	Synset element (+WN)
put	put into a certain place or abstract location	
		insert
		place
		position
		put
		replace
		set
	be in settled place	
		place down
		put down
	hold back for later	
		hold over
		put off
		put over
		set back
	set up for use	
		come in
		install

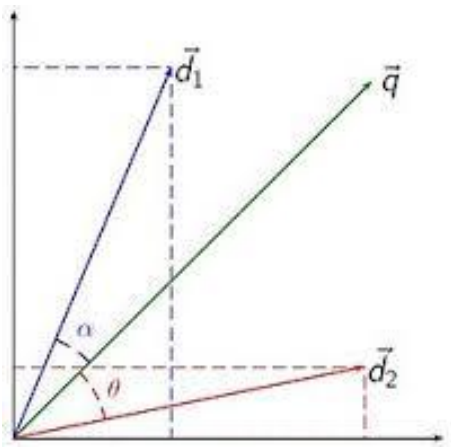
VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Experiment – synset growing ("add")

Initial Action	Synset	Synset element (+WN)	Synset element (+VN)
add	make an addition (to)		
		add	
			combine
			join
			mix
	add to something		
		add	
		put on	
	add to the very end		
		append	
		add on	
	determine the sum of		
		add up	
		add	
		total	
		sum	
		sum up	
			count

Word Space Model (WSM)

- Word Space Model (Shutze, 1995) is based on hypothesis, that "**words with similar meanings will occur with similar neighbors** if enough text material is available" (Shutze and Pedersen, 1997).
- WSM is implemented by:
  - calculating feature vector (e.g. frequency of co-occurence with other words) for each word
  - measuring the distance between two vectors – cosine similarity method



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Word Space Model (cont.)

- For each of synset verbs *co-occurrence matrix* obtained using *pointwise mutual information (PMI)* statistical measurement method:

$$PMI(A,B) = \log\frac{p(A,B)}{p(A)p(B)} = \log\frac{p(A|B)}{p(A)} = \log\frac{p(B|A)}{p(B)}$$

  - here where *p(A,B)* is a probability of words A and B joint distribution and *p(A), p(B)* – their individual distributions.
- Values from *co-occurrence matrix* are used for building the *vector of verb sense* and comparing by *cosine similarity* method:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}},$$

  - where *A* and *B* are vectors of verb senses;
- *Cosine similarity* ranges from -1 to 1, where -1 means exactly opposite sense, 0 – independence senses and 1 – exactly synonyms

Experiment – measuring verb similarity

WORD	PMI (wash)	PMI (rinse)
acetone	6,11	7,329
acid	6,15	4,108
after	5,33	6,236
away	6,31	6,823
be	3,39	5,909
careful	7,204	7,374
chemical	5,11	5,705
cleaner	8,29	7,959
container	6,42	5,254
crystals	4,93	3,23
dilute	6,55	5,636
discard	11,749	8,58
dish	8,269	8,464
distilled	6,981	6,213
down	5,96	5,814
drain	7,26	7,897
dried	5,37	4,884
dry	5,31	4,969
ether	6,77	3,344
expect	6,54	7,11

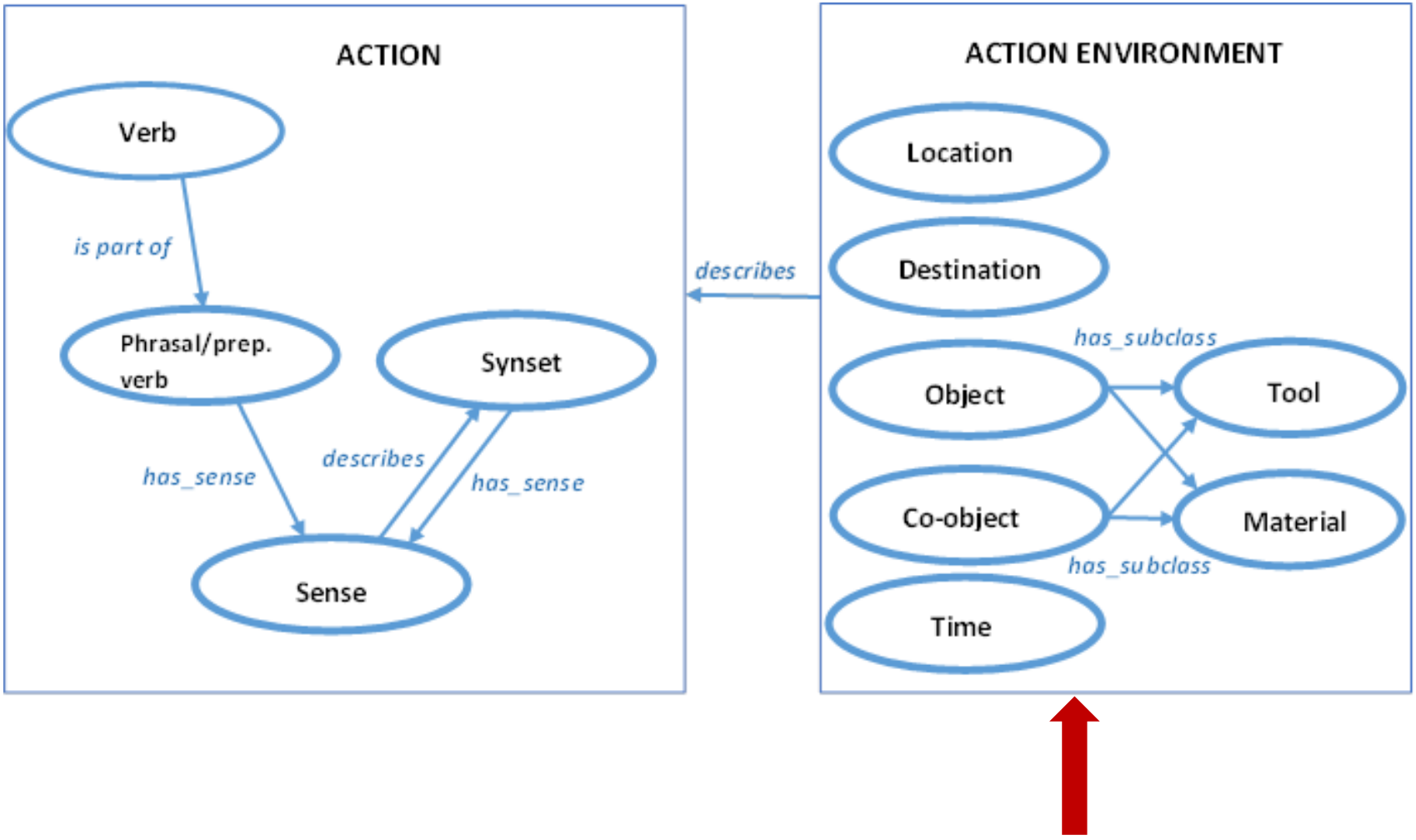
(extract)

WASH & RINSE

A*B	**2270,252**
||A||	**53,64**
||B||	**54,78**
cos(A,B)	**0,772615**

Terms are similar (closer to 1)

2nd stage – action environment building



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Action environment learning

- **Action environment learning** (for each action):
  - **Extracting search patterns** (rules) for extracting action environment elements from structured datasets (e.g. Verbnet)
  - Input: morphologically annotated corpus
- Another approach (or combination) - syntactic analysis can be used for extracting sentence structures, indicating thematic roles
  - Input: syntactically annotated corpus, parse trees

- **Text preprocessing –** collocation identification, named entity recognition
- **Harvesting action ontology** with action environment elements by applying search patterns

Text preprocessing – collocation identification

- **Collocation** – a sequence of words that co-occur more often than it would be by chance (e.g. "room temperature")
- **Different statistical methods** for collocation extraction – Mutual Information (MI), chi-squared test, log-likelihood ratio, Fisher's exact test, Dice coefficient, etc.
  - Usually applied in combination to rule-based techniques can be applied
- **Dice coefficient:**

$$D(A,B) = \log \frac{2|A \cap B|}{|A|+|B|},$$

  - here |*A and B*| is the frequency of *A* and *B* words co-occurrence in text, |*A*|, |*B*| - frequency of A and B words occurring separately.
  - logDice to fix very small numbers:

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

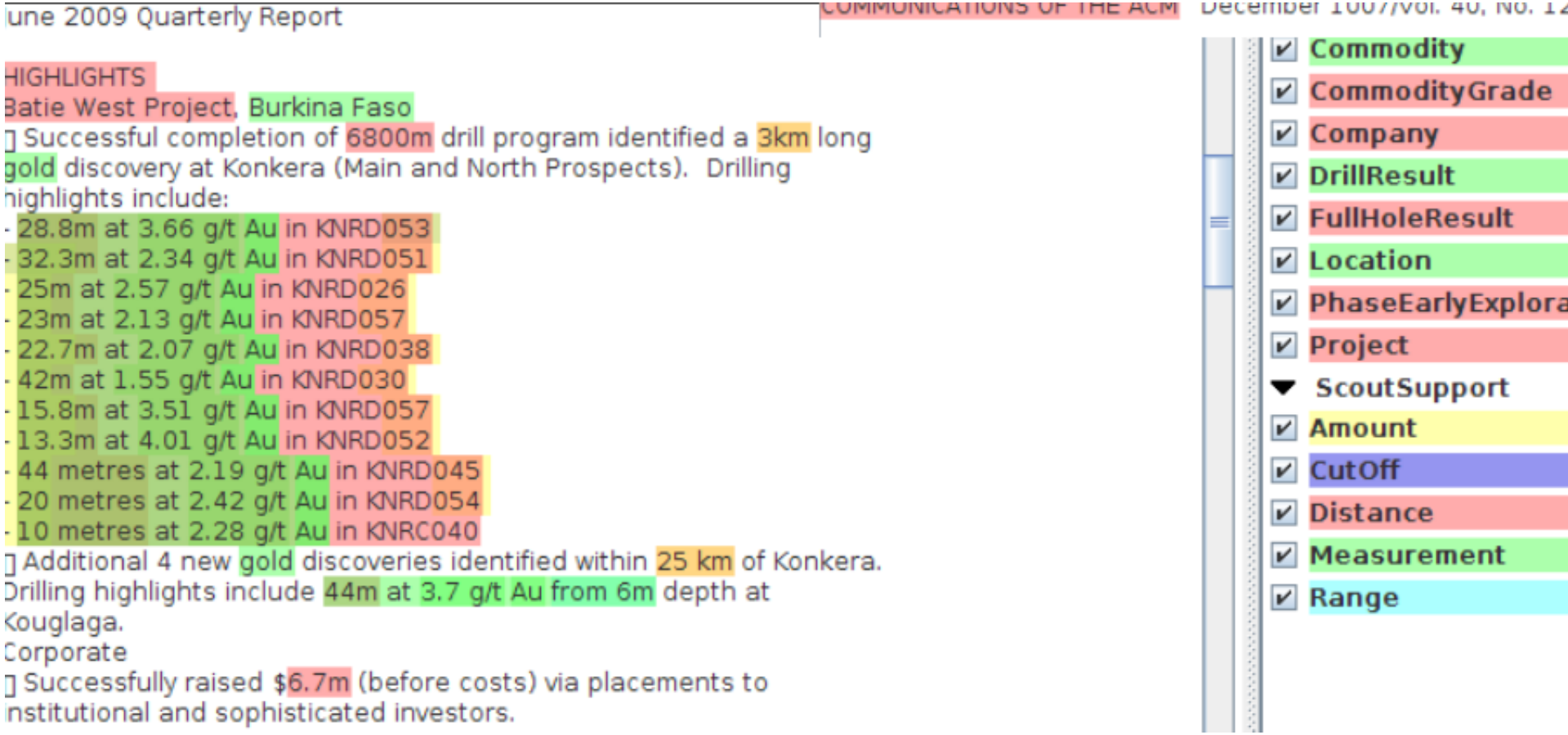
Text prepocessing – Named Entities Recognition (NER)

- Part of Information Extraction area
  - **identification** of proper names and **classification** into a set of predefines categories
  - Usually: person, location and organizaton
  - Other: date, **time**, **measures** (weight, money, …)
  - **Domain specific stuff**
  - **NOT event recognition & NO linking & NOT just matching text**

Named Entities Recognition (NER) - approaches

- **List lookup:**
  - Entities from gazetteers
  - Pros: simple, fast and easy to modify
  - Cons: collection and maintenance of lists, problems with variants, **cannot resolve ambiguity**
- **Rule-based:**
  - Context defined using rules
  - Regular Expressions (regexp)
  - JAPE (Java Annotations Pattern Engine, GATE)
    - Rules for manipulating annotations, regular expressions and entities lookup
- **Machine learning:**
  - Approaches
    - Supervised: CRF, HMM, SVM, …
    - Semi-supervised: bootstrap approaches, e.g. rules and then ML
    - Unsupervised: clustering
  - Pros: better learning of context (at least, potentially)
  - Cons: sufficient amount of training examples are necessary

Experiment – recognizing measurements and chemical elements



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Experiment: extracting VerbNet frames for action environment classification (part I)

Description	Syntax	Semantics	Example
NP V NP	NP-**Agent** VB NP-**Object**	**TAKE CARE OF:** ThemeRole = (?)Agent Event = during(E) ThemeRole = Object	***Wash*** *the aqueous* ***layer*** *twice.*
NP V	NP-**Agent** VB	**TAKE CARE OF:** ThemeRole = Agent Event = during(E) ThemeRole = (?)Object	*He* ***washed*** *the solvent* ***layer****, dried it and concentrated.*
NP V NP PP.instrument	NP-**Agent** VB NP-**Object** PREP-**With** NP-**Instrument**	**TAKE CARE OF:** ThemeRole = (?)Agent Event = during(E) ThemeRole = Object **USE:** ThemeRole = Agent Event = during(E) ThemeRole = Instrument	*The filter cake is* ***washed*** *thoroughly with* ***methanol.***

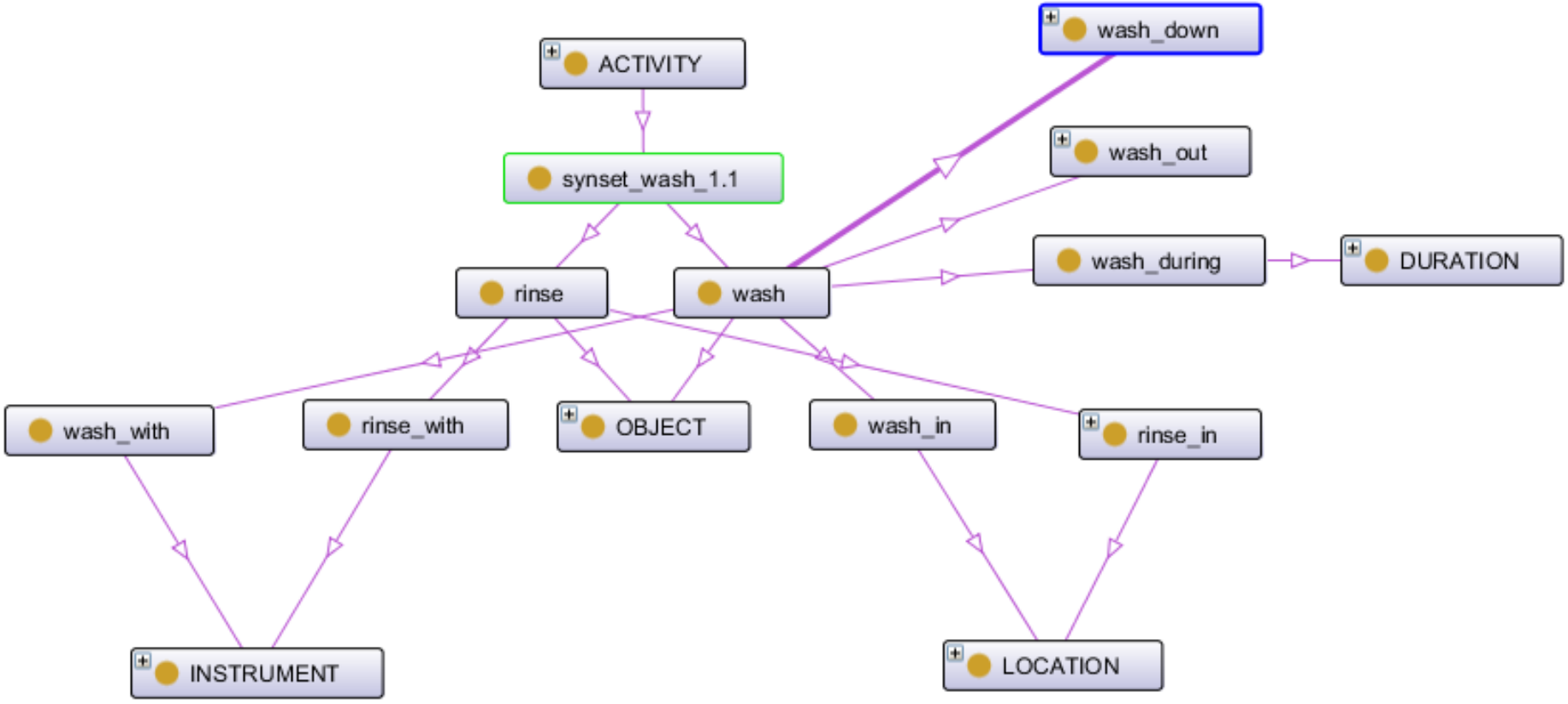
VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Experiment: extracting VerbNet frames for action environment classification (part II)

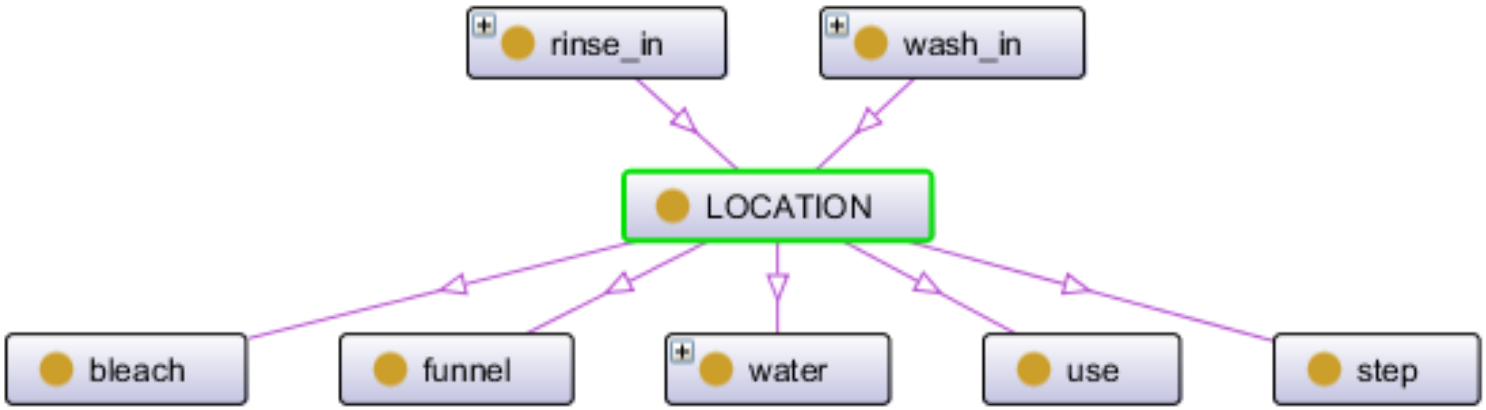
Description	Syntax	Semantics	Examle
NP V NP PP.location	NP-**Agent** VB NP-**Object** PREP-**In** NP-**Location**	**TAKE CARE OF:** ThemeRole = (?)Agent Event = during(E) ThemeRole = (?)Object **USE:** ThemeRole = Agent Event = during(E) ThemeRole = Location	*The top aqueous layer was **washed** in the **funnel.***
NP V NP PP.duration	NP-**Agent** VB NP-**Object** PREP-**During** NP-**Duration**	**TAKE CARE OF:** ThemeRole = Agent Event = during(E) ThemeRole = (?)Object **USE:** ThemeRole = Agent Event = during(E) ThemeRole = Duration	*The successive **washes** during the **work** up.*

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Excerpt of an experimental ontology (I)

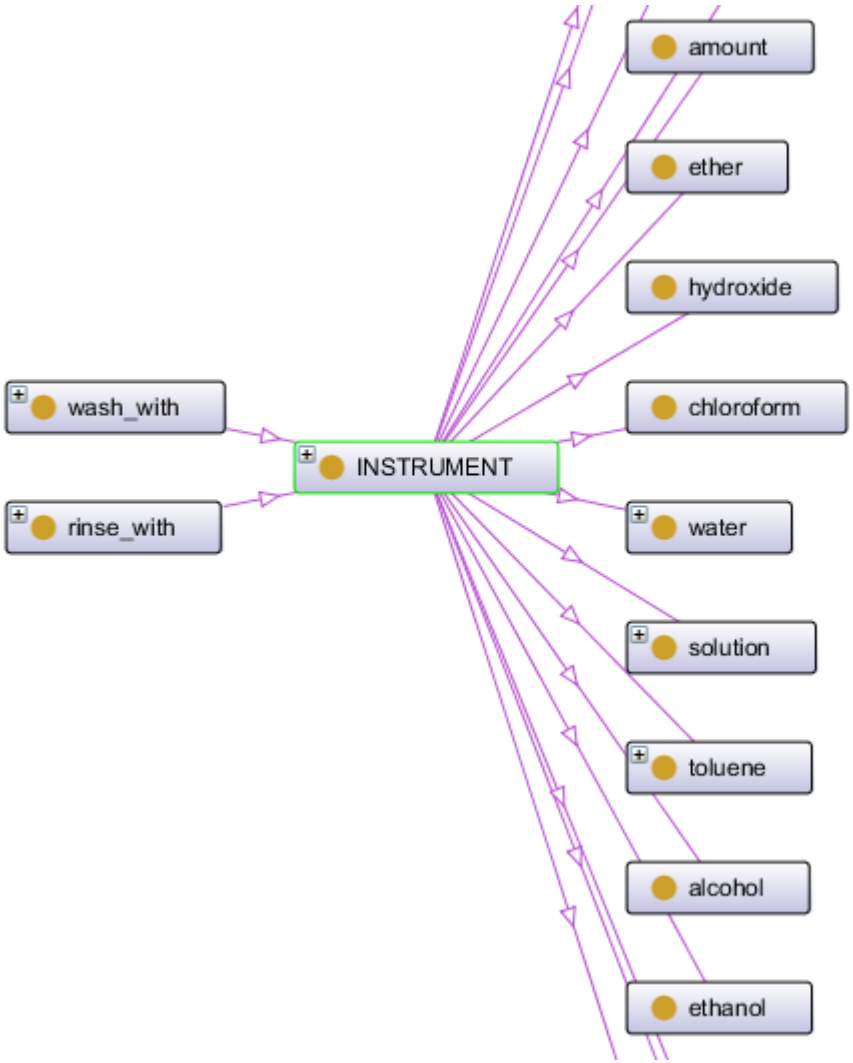


Excerpt of an experimental ontology (II)



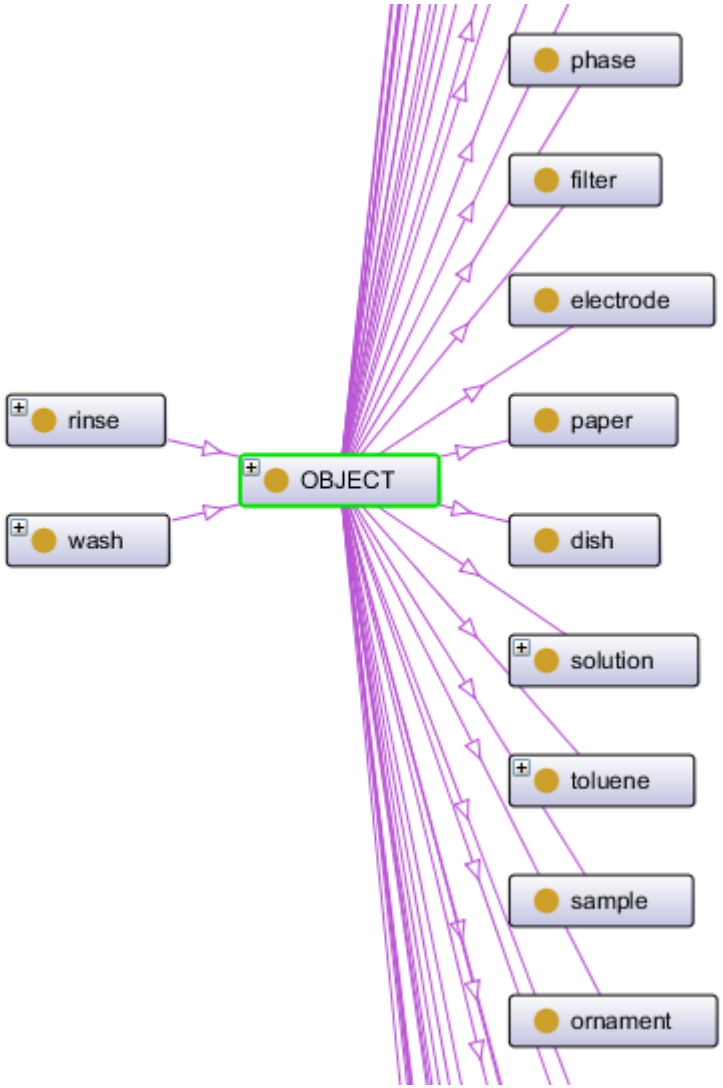
VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Excerpt of an experimental ontology (III)



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Excerpt of an experimental ontology (IV)



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Next steps

- Recognizing hierarchical relationships
  - Trained classifiers
  - E.g. built mapping Wordnet relations to morphosyntactically annotated corpora
- Building additional semantic relationships
  - Manually-built rules
  - Domain-specific rules obtained from other sources

Summary and conclusions

- Unsupervised and supervised ontology learning methods compensate each other's pros and cons and their combination gives better results in ontology building

- Structured information from existing knowledge bases (Wordnet, Verbnet, etc.) can be of use in designing automated procedures both for ontology concept and relation learning

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Thank you!

Contacts:

d.vitkute@if.vdu.lt

i.markievicz@if.vdu.lt

t.krilavicius@if.vdu.lt