DATA SCIENCE MASTER

SEMANTIC KNOWLEDGE REPRESENTATION

Ontology Learning and data mining

Mounira Harzallah

mounira.harzallah@univ-nantes.fr

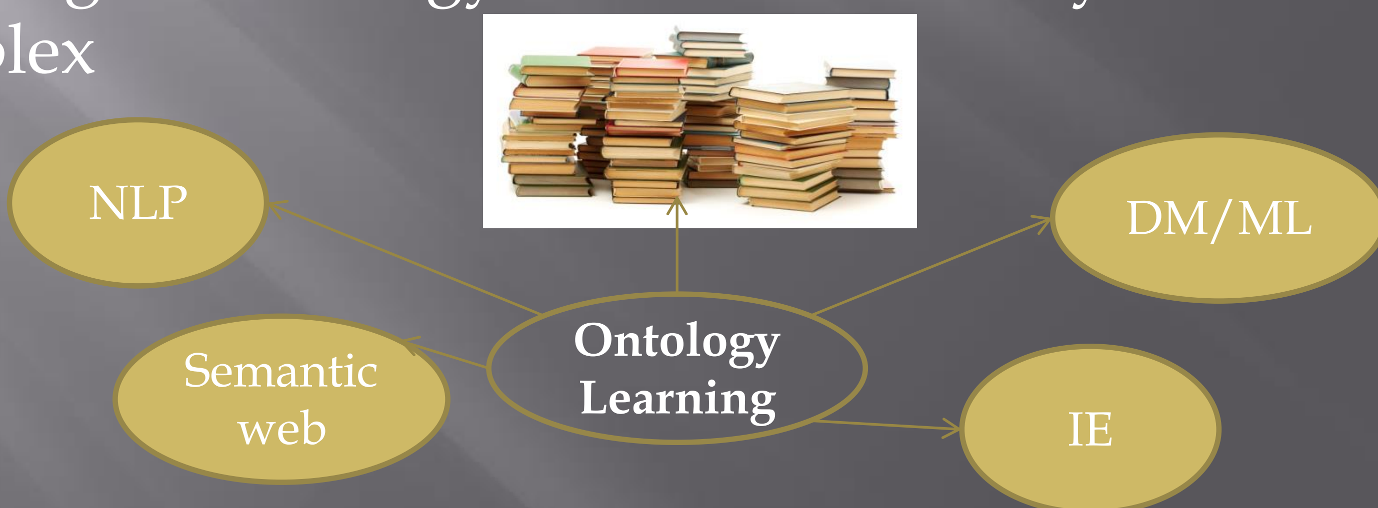
Tél: 02 28 09 21 27

Ontology learning from texts

Ontologies can be used in various domains

How can we have an ontology for a specific domain ?

- Resuse of existing ontologies: Ontology is not available for each domain.
- Building an ontology manually: knowledge acquisition bottleneck and time consuming
- Building an ontology semi-automatically fom texts : complex



Ontology learning from texts

Ontology conceptualisation and texts

Texts are interesting ressources, if the expert availability is limited or/and if the domain is large and complex

Texts:

- Content knowledge defined in natural langage
- May cover a domain well
- Content more or less accepted knowldge (famous books or articles)

Ontology learning from texts

Ontology conceptualisation and texts

Textual level → Conceptual level

Non-structured Knowledge → Structured Knowledge

How ???

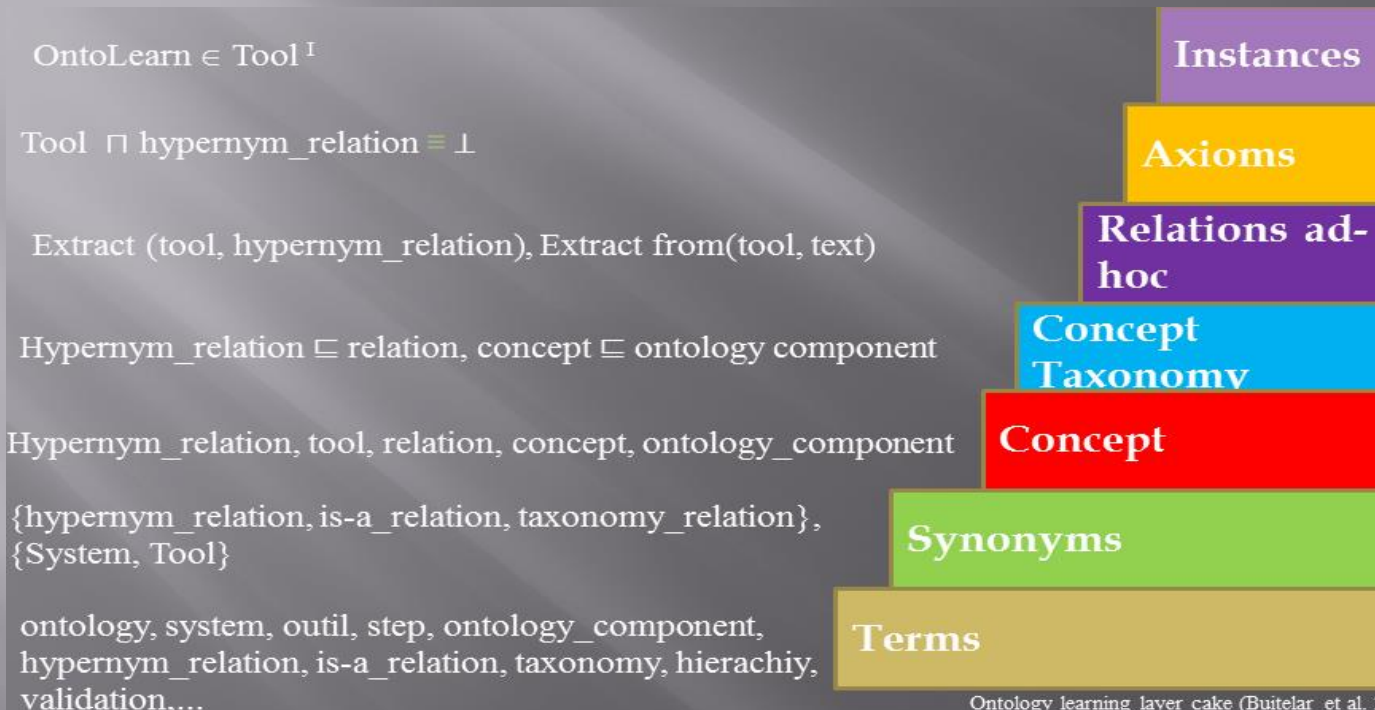
Manually is time consuming and complex if a huge number of texts Is considered

Semi-Automatically using

- NLP techniques
- Statistics, Data mining
- Machine learning

Ontology learning from texts

Techniques and tools are availables for each step



Ontology learning layer cake (Buitelar et al. 2

- ❑ No well recognized techniques for axioms extraction
- ❑ Selection of corpora about the domain of the ontology to build
- ❑ First version of the ontology to improve
- ❑ Need of experts for the validation of the tool results

Ontology learning from texts

Semi-automatic ontology building

Use of external resources for the identification of relations between terms.

e.g. Wordnet to identify synonym terms

Wordnet indicates that Car and Autocar are synonyms

Ontology learning from texts

Natural Language Processing (NLP) allows to extract **terms from a corpus**

→ **candidate terms for the ontology to build**

- Statistical techniques
- Linguistic techniques

Two approaches for relation extraction

- NLP + patterns matching to identify relations between terms
  - Composed terms
  - Synonym relations
  - Hypernym relations
  - Ad-hoc relations

- NLP+Distributional approaches for relation identification
  - Terms that co-occure frequently togother are related
  - Terms that occur frequently in similar contexts are semantically close

Ontology learning from texts

NLP. Text preprocessing

1. Lexical analysis:
   - Sentence splitting/segmentation : the process of dividing a text into sentences
   - Tokenising : to determine lexical units/words (token) of a text.

2. Part-of speech (POS) tagging : determine the part of speech for each word of a sentence = associate a grammatical tag to each word (e.g. Noun, Verb, Adjective (see the file « Tagg with NLT » in madoc) and its morphological characteristics (feminine, masculine, singular, plural, ,..)

3. Lemmatising : the process of grouping together the inflected forms of a word so they can be analysed as a single item, called lemma

- **Woman, Women → Woman**
- **Requires → Require**
- **The meeting → meeting**
- **We are meeting in this room → meet**

Ontology learning from texts

NLP. Text preprocessing

4. Syntactic analysis **: determination of** relations between lexical units/terms/words.
Example : Noun is a <u>subject</u> of a Verb
Adjectif is <u>a modifier</u> of a Noun

- Shallow parsing with heuristics + pattern matching
- Deep parsing with syntax trees

Ontology learning from texts

NLP. Text preprocessing

4. Syntactic analysis

Natural/language/ processing/tools/extract/terms/automatically / from /texts

noun, **verb**, **adjective** , **adverb**, verb-gerund, **preprosition**

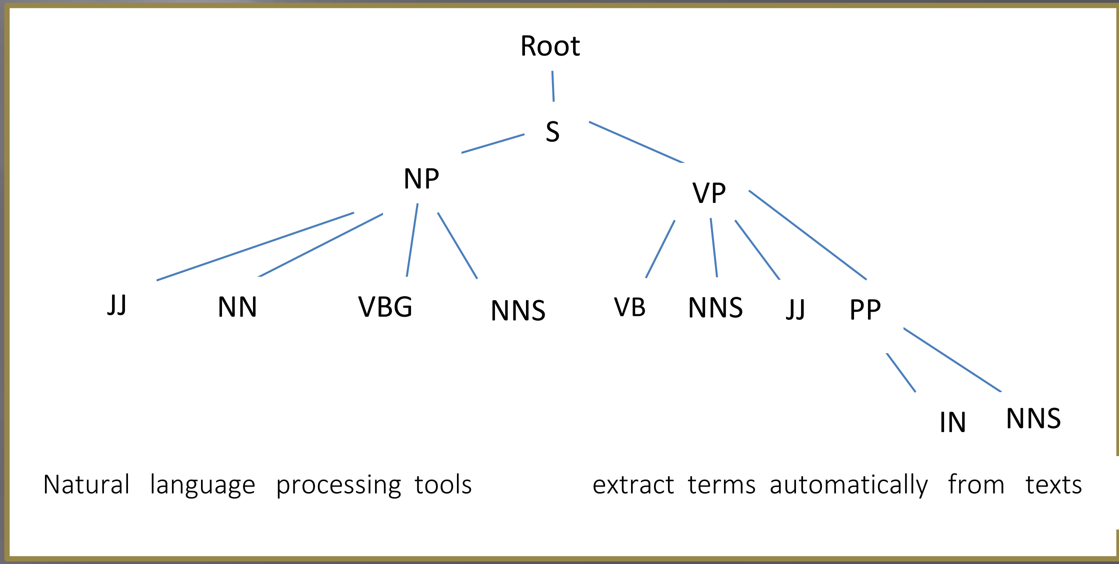
Shallow parsing :

Determine a Noun Phrase (NP) with the pattern JJ/NN/VBG/NN →NP

→ Natural language processing tool is a NP

NP/VB (Verb in the active voice) → NP is the subject of VB.

Deep parsing :



SubjectVerb (NLPT, extract from)
Object-Verb (Term, extract)
In_Object-Verb(Text, extract from)

Ontology learning from texts

NLP. Text preprocessing

4. Syntactic analysis. Deep parsing avec Stanford Parser

Please enter a sentence to be parsed:

```
Natural language processing tools extract
terms automatically from texts.
```

Language: English ▾   Sample Sentence   Parse

Your query

Natural language processing tools extract terms automatically fr

Tagging

```
Natural/JJ  language/NN  processing/NN  tools/NNS  extract/VB  terms/NNS  automatically/RB  from/IN  texts/NNS
```

Parse

```
(ROOT
  (S
    (NP (JJ Natural) (NN language) (NN processing) (NNS tools))
    (VP (VB extract)
      (NP (NNS terms))
      (ADVP (RB automatically))
      (PP (IN from)
        (NP (NNS texts))))
    (. .)))
```

Universal dependencies, enhanced

```
amod(tools-4, Natural-1)
compound(tools-4, language-2)
compound(tools-4, processing-3)
nsubj(extract-5, tools-4)
root(ROOT-0, extract-5)
dobj(extract-5, terms-6)
advmod(extract-5, automatically-7)
case(texts-9, from-8)
nmod:from(extract-5, texts-9)
```

Ontology learning from texts

NLP. Text preprocessing

Term filtering/selection : which terms to select for the ontology to build ?

Linguistic filtering/selection of terms

- Terms tagged with NP can be considered as terms of the ontology to build
- Terms tagged with Proper Noun (NNP) can be considered as instances/individuals of the ontology to build
- Terms tagged with verb can be considered as relations of the ontology to build.

Ontology learning from texts

Term Filtering/Selection : Removing irrelevant terms

Statistical approach

- tf : term frequency
- tf idf: term frequency-inverse document frequency

$$tfidf(w) = tf \,.\log\left(\frac{N}{df(w)}\right)$$

→ candidate term/relevant term extraction from texts

NLP. Text preprocessing.

Tools

- Gate framework (POS taggers + pattern matching)
- NLP/tm (text mining) librairy of R
- Python NLTK : NLP with Python
- Stanford CORE NLP library
- Spacy
- Etc.

Ontology learning from texts

Pattern based approach for relation extraction

A relation pattern that matches a sentence allows to extract a couple of terms related by this relation.

- Synonym relation. Pattern: NP i.e NP ; NP means NP

  Subsumption relation i.e. hypernym relation

- Hypernym relation : Hearst's patterns

  "NP such as NP,NP, NP …. and NP" matches "ontology components such as concept, relation and instance → and allows to extract « concept », « relation » and « instance » are hyponyms of « ontology component »

  Hearst's patterns : sometimes false results

  e.g. chocolate is a big problem in the context of children's health

In general HPs have low recall and good precision

Structural technique for hypernym relation extraction : based on the structure of a term term is a spcialisation of its head

"Domain ontology" is a "Ontology"

Ontology learning from texts

Pattern matching for relation extraction

Natural language processing tools extract terms automatically from texts

- Ad-hoc relation

« **/NN/**/extract(VB)/**/«from»/**/NN » → *Extract From(NN, NN)* : ExtractFrom(tool, text)

« **/NN/**/VB/**/NN » → VB(NN, NN) : Extract(language, term), Extract(tool, term), Extract(language, text), Extract(tool, text)

Pattern matching weakness

- Law recall for the hypernym pattern
- Ad-hoc patterns
  - Depend of the domain of a corpus
  - Define manually patterns for each corpus is not easy

    → Pattern learning from corpus

Ontology learning from texts

Distributional based approach.

Harris' distributional hypothesis : terms (or pairs of terms) that occur in similar contexts tend to have similar meanings (or be related with the same relation)

- Co-location : terms occur <u>frequently</u> in similar contexts

T1C1*, *T2C1*, *T1C2*, *T2C2*…. → *T1* *and* *T2* *are close semantically

<u>Tool</u> extracts…. <u>System</u> extracts….. <u>Tool</u> learns….. <u>System</u> builds …..

It is extracted from <u>documents</u>. It is extracted from <u>texts</u>.
It is learned from <u>documents</u>. It is learned from <u>texts</u>

Ontology learning from texts

Distributional based approach.

Harris' distributional hypothesis : terms (or pairs of terms) that occur in similar contexts tend to have similar meanings (or be related with the same relation)

- Co-Occurrence : terms occur frequently together
  T1C1**T2**, **T2T1**C2, **T1**C3**T2**, … → ***T1*** *and* ***T2*** *are related semantically*

→**Identification of relations between terms**

T1C1**T2**, **T2T1**C2, **T1**C3**T2**, …
T3C1**T4**, **T3T4**C2, **T3**C3**T4**, …

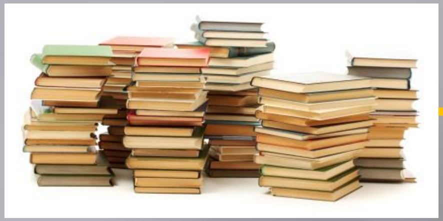
→ ***(T1,T2) and (T3 ,T4)*** *are related by the same relation*

Ontology is learned from texts.	Ontology is build from texts

Taxonomy is learned from documents.	Taxonomy is build from documents

Ontology learning from texts

Distributional based approach.



Verbe Sujet	Extract	learn	Is Extracted	compose	…
system	30	15	0	0	…
Tool	20	20	0	0	…
Hypernym	0	0	20	20	…
Ad_hoc R	0	0	10	10	…
Concept	0	0	20	25	…
OntoLearn	15	25	0	0	…
…	…	…	…	…	…

Matrix space model

or

System	Tool	Hypernym	Ad-hoc R
30	20	0	0
15	20	0	0
0	0	20	10
0	0	20	10
20	21	0	0
12	15	0	0
0	0	15	10
10	9	0	0
2	0	10	13

Vector space model

- Individuals : terms or pairs of terms to cluster or to classify
- Features : term or pair contexts, term tagging, etc, …..

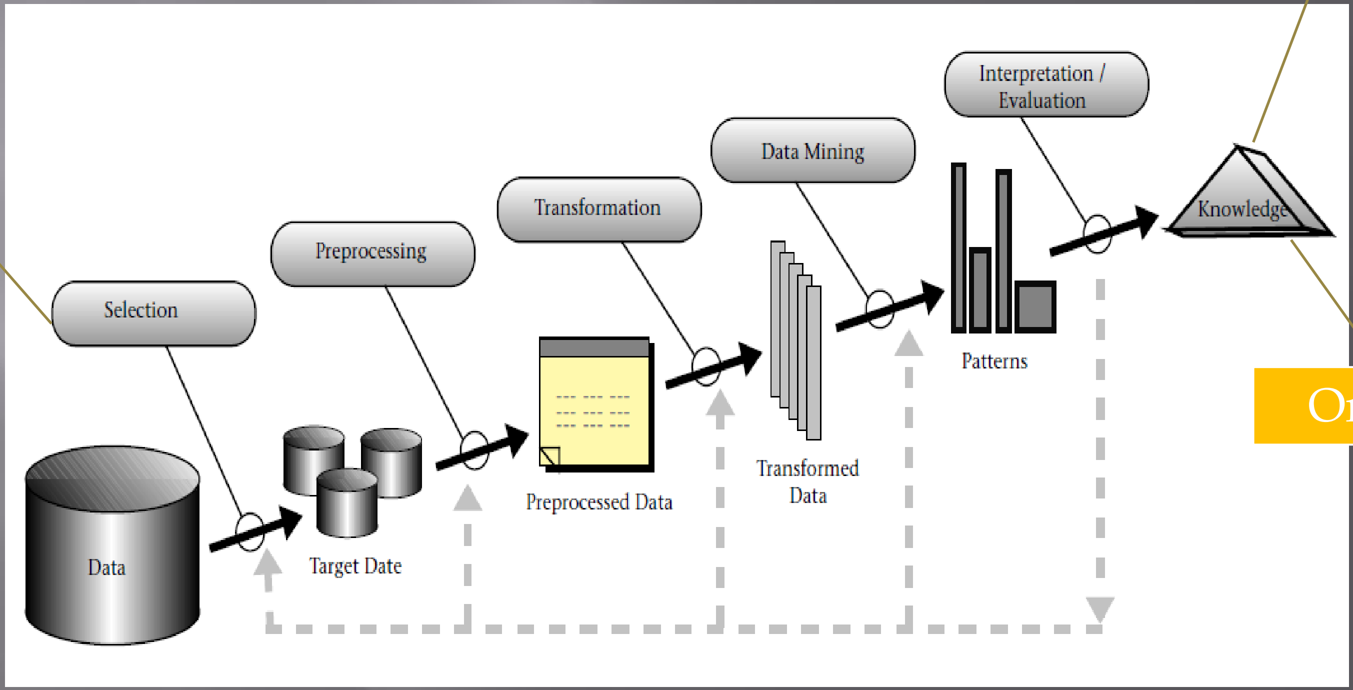
→apply data mining/ machine learning techniques following the KDD process

Ontology learning from texts

Data mining/machine learning approach.

Clusters, Association rules, Prediction, …

Structured data



Ontology

Texts

Steps of the process of KDD [Fayyad *et al.* 1996]

Ontology learning from texts

Data mining/machine learning approach.

Adaptation of one or many methods

+ Core ontology

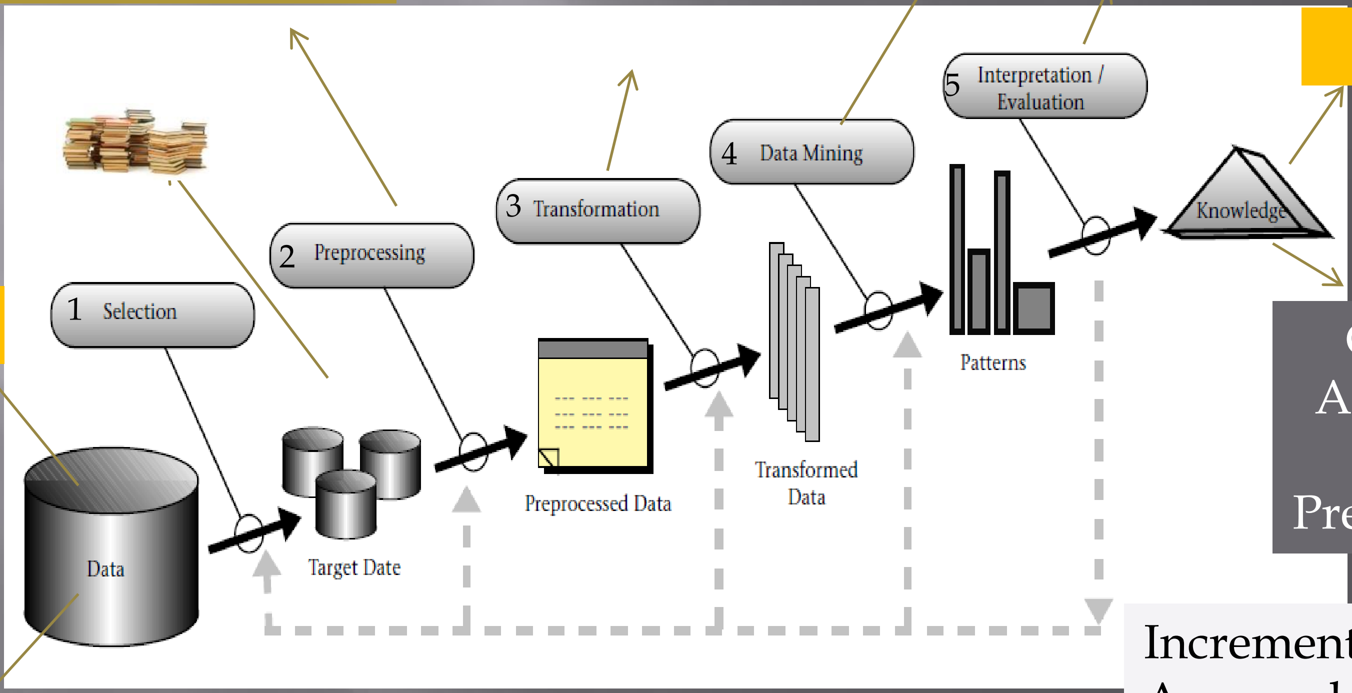
Individual and feature selection to prepare the matrix/vector model

Semi-automatic methods

Schallow or Deep Parsing

Ontology

Texts



Clusters, Association rules, Prediction, …

Structured data

Incremental Approach

Steps of the process of KDD [Fayyad *et al.* 1996]

Ontology learning from texts

Data mining/machine learning approach.

- **Clustering**
  - Cluster of terms semantically close → Concept of the ontology

    → Cluster associated to a core concept

Ontology learning from texts

Data mining/machine learning approach.

- Ontology
- Process
- Step
- Component
- Tool
- Resource
- User
- Technique

Resource
Corpora, Domain corpora
Web site
Document Warehouse

User
Domain specialist
Knowledge engineer

Ontology
Domain ontology
Specialised ontology
Computational ontlogy

Tool
Automatic System

Step
Evaluation
Ontology evaluation, concept evaluation
Natural language description
Formal specification
Formal concept specification
Concept qualitative analysis
Automatic acquisition
Word sens disambiguation

Component
Concept, learnd concept
Complex domain concept

Technique
Syntactic pattern matching technique
Ontology learning Algorithm

Ontology learning from texts

Data mining/machine learning approach.

- **Clustering**
  - Cluster of terms semant… the ontology
  - Cluster of pairs of terms…
- **Classification**
  - Classify a term into a…
  - Classify a pair of terms into class, corresponding to a predefined relation
    - Synonymy relations → from terms to concepts
    - Hyperonymy relations →taxonomy definition
    - Ad-hoc Relations →ontology definition

mots entre Couples	Especi ally	Such as	Extract from	Process	Includin g	Y	N
System, NLP tool	20	8	0	0	10	0	1
Component, Concept,	7	10	0	0	15	1	0
Relation, Hypernym,	15	10	0	0	10	1	0
NLP, website	0	0	25	10	0	0	1
Relation, Ad_hoc Re.	15	10	0	0	13		
OntoLearn, text	0	0	56	15	0		

Need of terms or pairs already classified

Ontology learning from texts

Steps to do

- Choice and selection of the contexts (discriminants)
- Dimensionality Reduction
- Choice of the clustering/classification technique
- Result interpretation

Steps for data preparing

Ontology learning from texts

Choice of the kind of contexts

Graphical context: term neighbors → a window arround the term

Natural language processing tools extract $\boxed{\text{terms}}$ automatically **from texts**

-3 +3

Syntaxic contexts

- Gramatical tagging of neighbors of terms
- Verbs for which a term is a subject
- Dependency relations of terms of a pair

Ontology learning from texts

Selection of relevant contexts for clustering/classification

- Use of a lexicon (stop words)
- Selection of contexts regarding their frequency
- Selection of contexts regarding their frequency with termes
  - Frequency (context/term)
  - TF-IDF(context/term)

To remove

→general contexts

→ contexts occur with all terms or occur with very few terms

Data mining/machine learning approach

- Matrix/vector models: Subject/Verb ; Object/verb

Verbe / Sujet	Extract	Learn	Is extracted	Compose	… …
system	30	15	0	0	…
Tool	20	20	0	0	…
Hypernym	0	0	20	20	…
Ad_hoc Re	0	0	10	10	…
Concept	0	0	20	25	…
OntoLearn	15	25	0	0	…
…	…	…	…	…	…

- Matrix terms/sentences

Phrase / NP	S1	S2	S3	S4	
Text2Onto	1	1	0	0	…
Tool	0	0	0	1	…
Hypernym	1	1	0	1	…
Ad_hoc Re	0	0	1	1	…
Concept	0	1	0	1	…
OntoLearn	0	0	1	0	…
…	…	…	…	…	…

Ontology learning from texts

- Matrix of terms and their neighbors in a window with a given a size

Sujet	Extract	Learn	Is_extracted	Compose	from	Web site	ressource	Ontology
system	30	15	0	0	20	10	10	0
Tool	20	20	0	0	20	15	10	0
Hypernym	0	0	20	20	0	0	0	10
Ad_hoc Re	0	0	10	10	0	0	0	25
Concept	0	0	20	25	0	0	0	10
OntoLearn	15	25	0	0	13	14	16	0
…	…	…	…	…	…			

Ontology learning from texts

- Matrix couples of terms/words between them

Liens Couples	Especi ally	Such as	Extract from	Process	Including	…
System, NLP tool	20	8	0	0	10	…
Component, Concept,	7	10	0	0	15	…
Relation, Hypernym,	15	10	0	0	10	…
NLP, website	0	0	25	10	0	…
OntoLearn, text	0	0	56	15	0	…
Relation, Ad_hoc Re.	15	10	0	0	13	…

Clustering/classificat iondes of couples relied by similar relations

Difficulties

- **Selection of relevant individuals and their features .**
- **Matrix Sparsity :**
  - **Matrix dimensionality reduction technique is required**
  - **Selection of a method adapted to a sparse matrix**

Ontology learning from texts

Data mining/machine learning approach.

- Matrix/vector model: pairs of terms/words between them

Liens Couples	Especi ally	Such as	Extract from	Process	Including	…
System, NLP tool	20	8	0	0	10	…
Component, Concept,	7	10	0	0	15	…
Relation, Hypernym,	15	10	0	0	10	…
Relation, Ad_hoc Re.	15	10	0	0	13	…
NLP, website	0	0	25	10	0	…
OntoLearn, text	0	0	56	15	0	…

Difficulties

- **Selection of relevant individuals and their features .**
- **Matrix Sparsity :**
  - **Matrix dimensionality reduction technique is required**
  - **Selection of a method adapted to a sparse matrix**

Ontology learning from texts

Data mining/machine learning approach.

Non supervised methods : partition methods, agglomerative clustering (CHA), Topic modelling (LDA),..

Verbe Sujet	Extract	Learn	Is extracted	Compose	… …
system	30	15	0	0	…
Tool	20	20	0	0	…
Hypernym	0	0	20	20	…
Ad_hoc Re	0	0	10	10	…
Concept	0	0	20	25	…
OntoLearn	15	25	0	0	…
…	…	…	…	…	…

Clusters/Hierachy of clusters of terms (pairs of terms)

- ❑ Selection of a similarity measure
- ❑ Difficulty to interpret the semantic of each cluster. Unlabled cluster
- ❑ Clustering is a first iteration that can be improved with others methods.

Ontology learning from texts

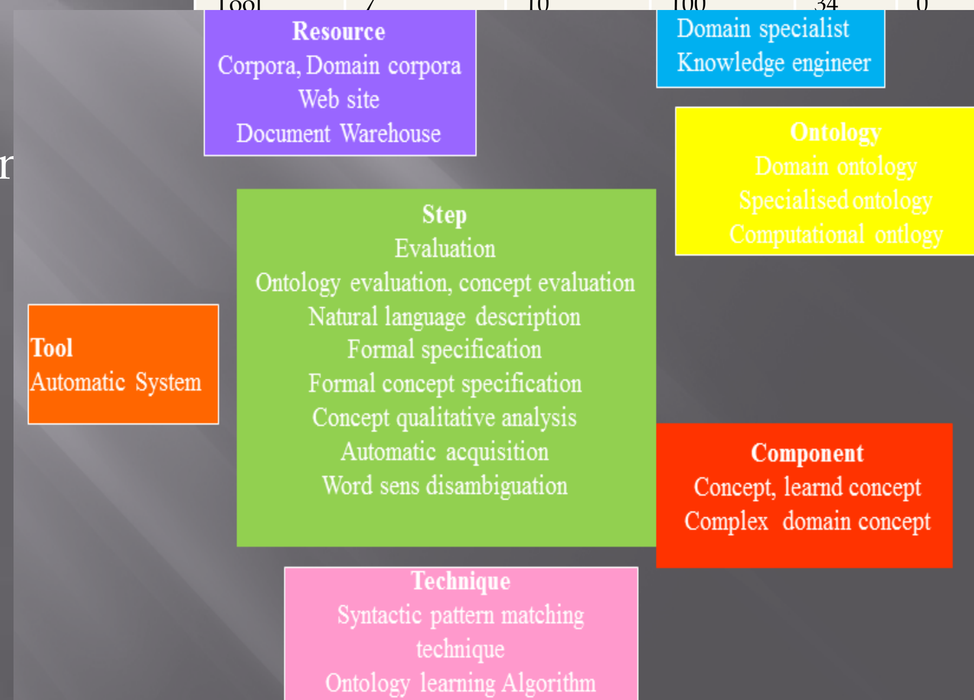
Non supervised method : K-means

- **Goal** : form clusters of terms (pairs of terms) semantically close : synonym terms terms that have the same hypernym, (pairs related by the same kinds of relation)

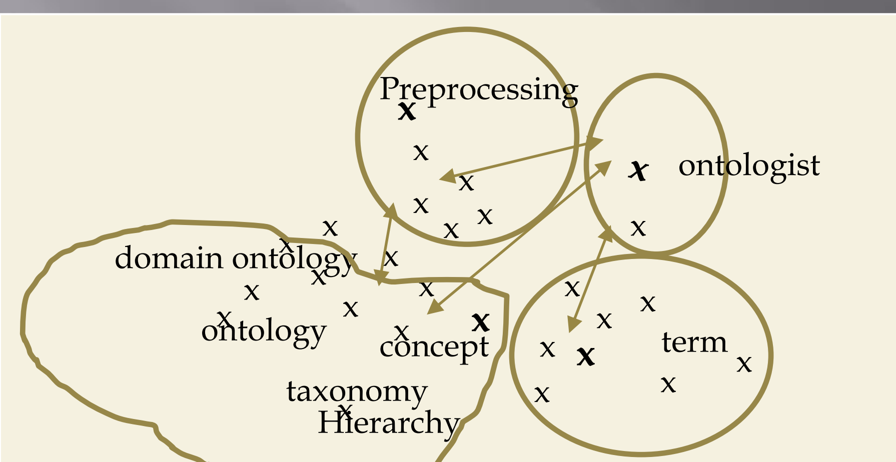
- **Difficulties** : Choice of K, Interpretation of classes , Choice of the distance

- **Result evaluation :**
  - Quality of the clustering (inter-cluster/ intra cluster distances) Inertie Intra-classe)
  - Precision and recall regarding a gold star each cluster can be compared to a class of hyponyms of a core concept

	Extract	learn	Extract from	Use	comp
system	20	8	45	17	0
Tool	7	10	100	34	0



text

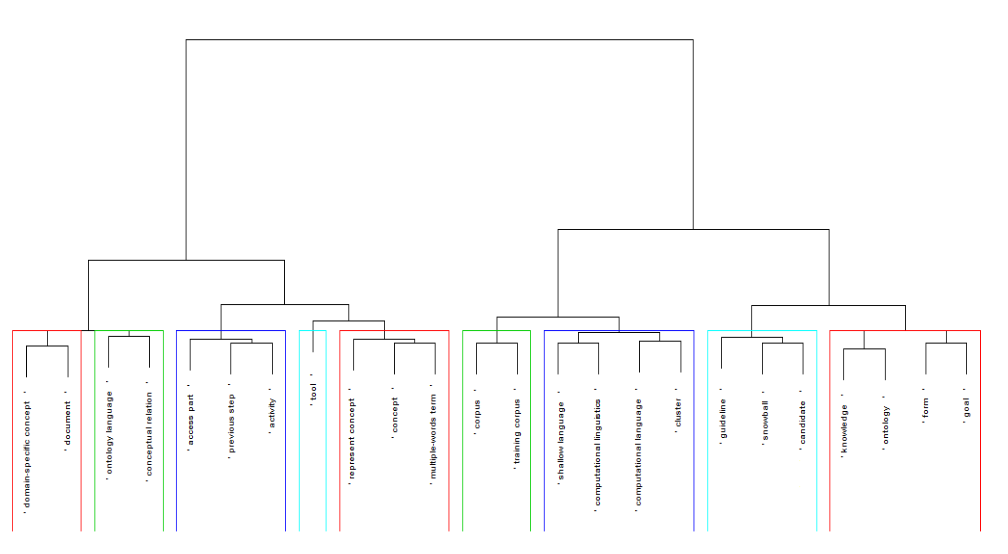


Ontology learning from texts

Non supervised method: Agglomerative Clustering method

	Extract	learn	Extract from	Use	compose
system	20	8	45	17	0
Tool	7	10	100	34	0
hypernym	0	0	0	0	10
Ad_hoc R	0	0	0	0	8
document	0	0	0	0	0

- **Goal :** form a hierarchy of classes of terms close semantically
  - Synonym terms , terms having the same hypernym
  - Hierarchy relation : isa, part-of relations
- **Difficulties and Evaluation (see k-means)**

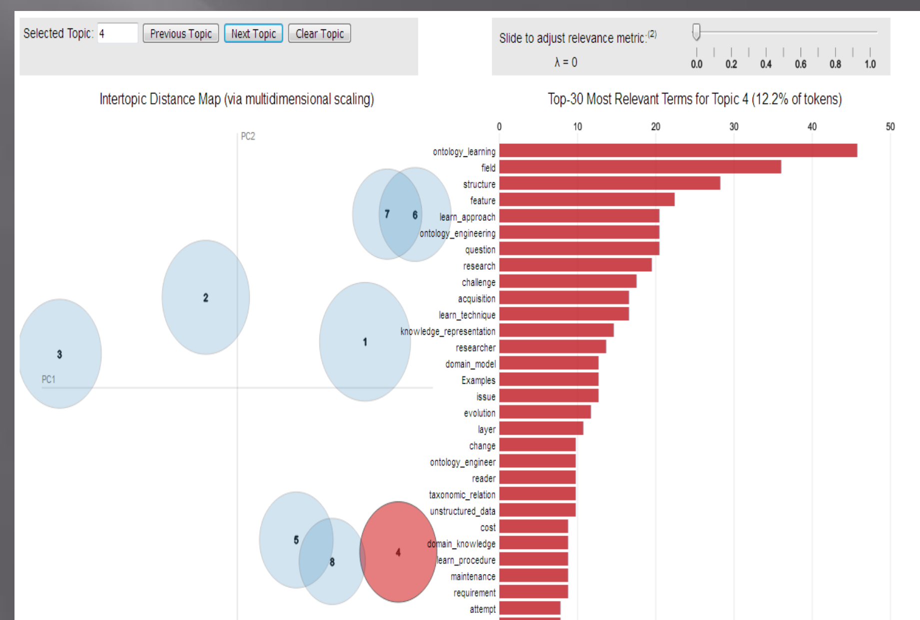


Ontology learning from texts

Méthodes non supervisées : Méthode LDA (Latent Distribution Analysis)

- **Objectif** : form clusters, each one includes terms dealing with the same topic (method adapted to sparcy matrix)
- **Modelisation**: matrix of documents/ terms

- **Difficulty:** method parameter choice, interpretation of each cluster and overlapping of topics.



Ontology learning from texts

Supervised method : KNN (k- nearest neighbors), SVM, neural network, decision tree, Naive Bayes Classifyer

- Goal: Associate a term or a pair of terms to a known/predefined class

- Modelisation
  - Matrix/vector model
  - Predefined classes
  - Training data set : labeled instances : a set of instances where each one is associated to a class

Verbe / Sujet	Extr[illegible]	Learn	Is extracted	Compose	Tool	Component
System	30	15	0	0	1	0
Text2Onto	20	20	0	0	1	0
Hypernym	0	0	20	20	0	1
Ad_hoc Re	0	0	10	10	0	1
Concept	0	0	20	25	0	1
OntoLearn	15	25	0	0	1	0
…						

Difficulties/weakness

- Individual and feature selections
- Classes should be previously defined : classes can correspond each to a core concept class
- Training data set (labeled instances) is required

Liens / Couples	Especially	Such as	Extract from	Process	Including	Y	N
System, NLP tool	20	8	0	0	10	0	1
Component, Concept,	7	10	0	0	15	1	0
Relation, Hypernym,	15	10	0	0	10	1	0
Relation, Ad_hoc Re.	15	10	0	0	13	1	0
NLP, website	0	0	25	10	0	0	1
OntoLearn, text	0	0	56	15	0	0	1

Ontology learning from texts

Supervised method : KNN

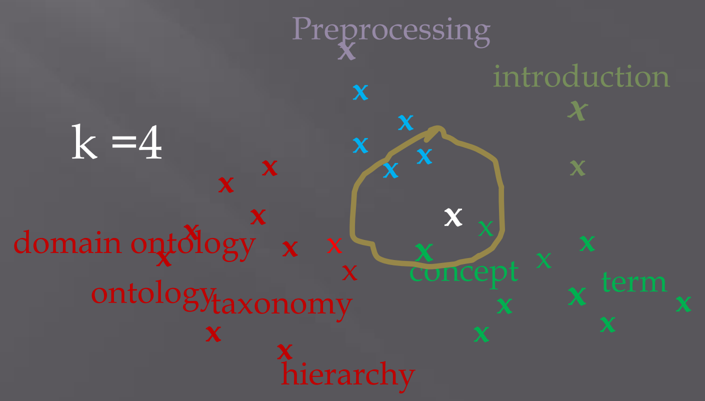
Objectif : classify a term regarding the k terms that are the most close to eat.

	Extract	lear n	Extract from	Use	comp ose	Is - built	1	2	3	4
syste m	20	8	45	17	0	0	0	1		0
Tool	7	10	100	34	0	0	0	1		0
term	0	0	0	0	10	0	1	0	0	0
conce pt	0	0	0	0	8	0	1	0	0	0
taxon omy	0	0	0	0	0	20	0	0	1	0
Introd uction	0	0	0	0	0	0	0	0	0	1

Difficulties : Choice of the method parameters (k), similarity measure, training data set

Result evaluation

- ❑Recall and precision



Conclusion

- NLP/Data mining/ ML are interesting for ontology learning

- Difficulties
  - Text selection and Filtration
  - Selection and adaptation of a method to a kind of a corpus
  - Identifications of relevant features
  - Matrix dimensionality reduction
  - Choice of method parameters
  - Supervised method : training data set definition
  - Non Supervised method : Interpretation of the results (classes) and their improvement

Combine several methods