AI and Law
Semantic Annotation of Legal Texts

Enrico Francesconi

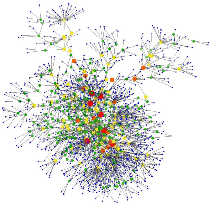
Publications Office of the EU
enrico.francesconi@publications.europa.eu

ITTIG-CNR – Institute of Legal Information Theory and Techniques
Italian National Research Council
enrico.francesconi@ittig.cnr.it

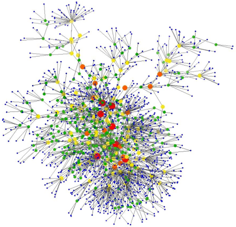
Central South University, Changsha – 16 April 2019

Semantic Annotation Approaches

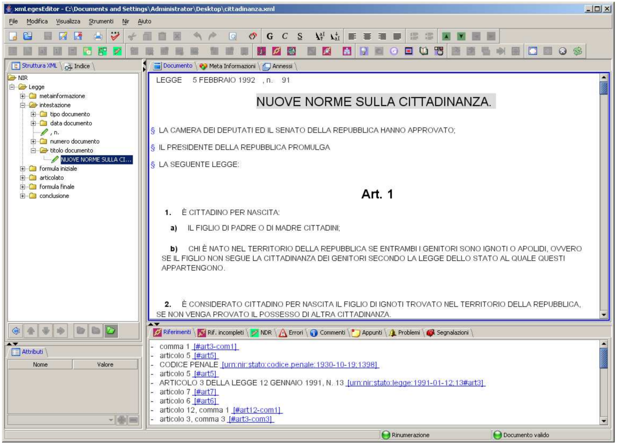
1. Bottom-Up semantic annotation
   - Manual
     - Editing environment for Provision Model semantic annotation
   - Automatic (semi-automatic)
     - Automatic Classification of Provisions (ML [Francesconi and Passerini, 2007], NLP [de Maat et al., 2010])
     - Provision Attributes Extraction (NLP [Biagioli et al., 2005])
2. Top-Down semantic annotation
   - Visual environment using the Provision Model as semantic guide for planning a new bill
3. Semantic interoperability
   - Mapping between knowledge models concepts



Semantic Annotation
Bottom-Up Approach

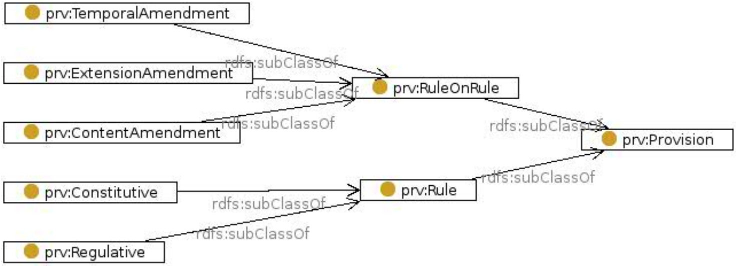


Legislative drafting environment

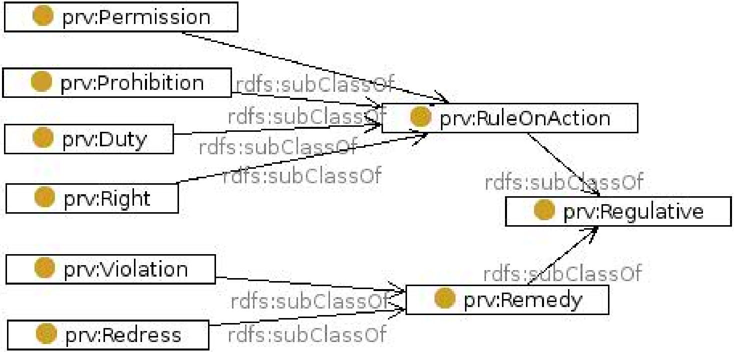


- URI and XML standards implementation
- Facilities for semantic annotation

Provision Model Top Classes



Regulatives provisions



Excerpt of EU Directive 2002/65/EC

Art. 5
1. The supplier shall communicate to the consumer all the contractual terms and conditions and the information referred to in Article 3(1) and Article 4 [...]

2. The supplier shall fulfil his obligation under paragraph 1 immediately after the conclusion of the contract, if the contract has been concluded at the consumer's request using a means of distance communication which does not enable providing the contractual terms [...]

3. At any time during the contractual relationship the consumer is entitled, at his request, to receive the contractual terms and conditions on paper. [...]

[...]
Art. 6
1. The Member States shall ensure that the consumer shall have a period of 14 calendar days to withdraw from the contract without penalty and without giving any reason [...]

[...]

Formal Profile: Set of paragraphs

Art. 5	
1. The supplier shall communicate to the consumer all the contractual terms and conditions and the information referred to in Article 3(1) and Article 4 [...]	Paragraph
2. The supplier shall fulfil his obligation under paragraph 1 immediately after the conclusion of the contract, if the contract has been concluded at the consumer's request using a means of distance communication which does not enable providing the contractual terms [...]	Paragraph
3. At any time during the contractual relationship the consumer is entitled, at his request, to receive the contractual terms and conditions on paper. [...]	Paragraph
[...]	
Art. 6	
1. The Member States shall ensure that the consumer shall have a period of 14 calendar days to withdraw from the contract without penalty and without giving any reason [...]	Paragraph
[...]	

Semantic Profile: Set of Provisions

Art. 5

1. The supplier shall communicate to the consumer all the contractual terms and conditions and the information referred to in Article 3(1) and Article 4 [...]

Duty (*Supplier*, *Consumer*)

2. The supplier shall fulfil his obligation under paragraph 1 immediately after the conclusion of the contract, if the contract has been concluded at the consumer's request using a means of distance communication which does not enable providing the contractual terms [...]

Procedure (*Supplier*, *Consumer*)

3. At any time during the contractual relationship the consumer is entitled, at his request, to receive the contractual terms and conditions on paper. [...]

Right (*Consumer*, *Supplier*)

[...]

Art. 6

1. The Member States shall ensure that the consumer shall have a period of 14 calendar days to withdraw from the contract without penalty and without giving any reason [...]

Duty (*Member States*, *Consumer*)

[...]

Automatic Classification of Provisions

Classifying paragraph according to provision types is a problem of document categorization

Two machine learning approaches of text categorization have been tested

- Naïve Bayes
- Support Vector Machine

Document Representation

A document is represented by a vector of term weights $d_j = (w_1, ..., w_{|T|})$ and three different types of weights have been tested:

- Binary weights (presence/absence);
- Term frequency weight (tf);
- TF-IDF weight (which penalizes terms occuring in many different documents, being less disciminative);

Pre-processing to increase statistical qualities of terms:

- Stemming (reduction of terms to their morphological root)
- Stopwords elimination (deletion of very frequent terms)
- Digits and non alphanumeric characters represented by a unique special character

Feature Selection

Terms Selection by

- an unsupervised min frequency threshold aiming at eliminating terms with poor statistics;
- a supervised threshold over the Information Gain of terms (discriminative power of a term with respect to the classes)

$$ig(w) = H(D) - \frac{|D_w|}{|D|} H(D_w) - \frac{|D_{\bar{w}}|}{|D|} H(D_{\bar{w}})$$

  - Information Gain in terms of Entropy ($H(D)$) reduction
  - Optimal case:
    given a word and a class if all the documents containing that word belong to that class $\Longrightarrow H(D_w) = 0$

    where $H(D) = \sum_{i=1}^{|C|} -p_i \; log_2(p_i)$

The Experiments

Data set of 582 examples (fragments of text containing a provision), belonging to 11 classes

Class labels	Provision Types	Number of documents
c_0	Repeal	70
c_1	Definition	10
c_2	Delegation	39
c_3	Delegification	4
c_4	Duty	13
c_5	Exception	18
c_6	Inserting	121
c_7	Prohibition	59
c_8	Permission	15
c_9	Penalty	122
c_{10}	Substitution	111

Naïve Bayes

Using paragraphs full text

Train Accuracy	LOO Accuracy	N terms with max InfoGain
90.7%	86.9%	100
89.3%	86.9%	50

Excluding quoted text ("misleading text")

Train Accuracy	LOO Accuracy	N terms with max InfoGain
95.5%	88.6%	500
94.3%	88.1%	250

SVM

Using paragraphs full text

Train Accuracy	LOO Accuracy	N terms with max InfoGain
100%	91.2%	1000
100%	91.9%	500

Excluding quoted text ("misleading text")

Train Accuracy	LOO Accuracy	N terms with max InfoGain
99.8%	92.1%	all
99.8%	92.1%	1000

Chunking and SVM

Text representation using linguistic structures of higher level of abstraction

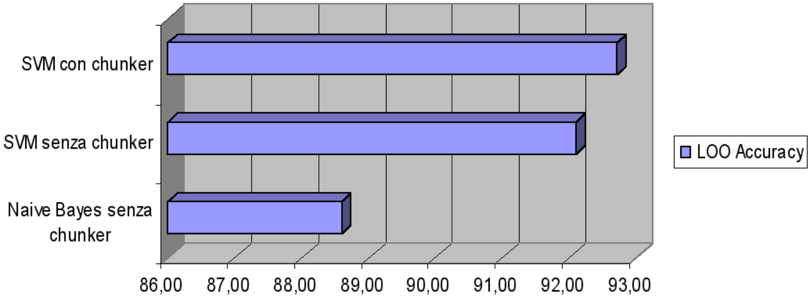
Using paragraphs full text

Train Accuracy	LOO Accuracy	N terms with max InfoGain
99.7%	92.4%	all
99.7%	92.4%	100

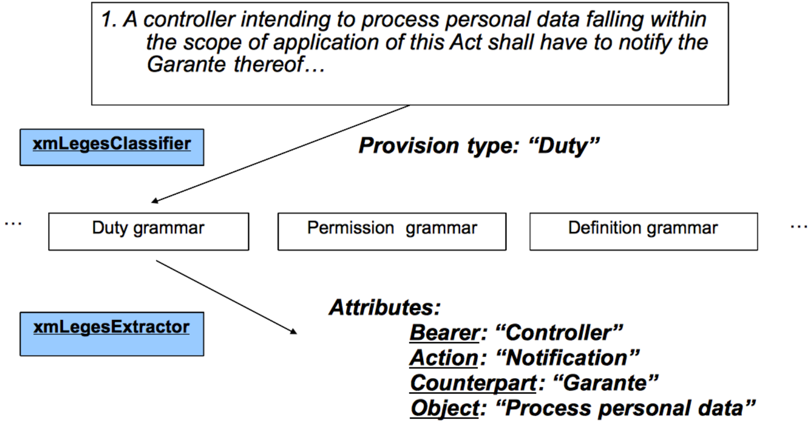
Excluding quoted text ("misleading text")

Train Accuracy	LOO Accuracy	N terms with max InfoGain
99.7%	92.7%	all
99.7%	92.7%	500

Comparison of the Results



Provision Attributes Extraction

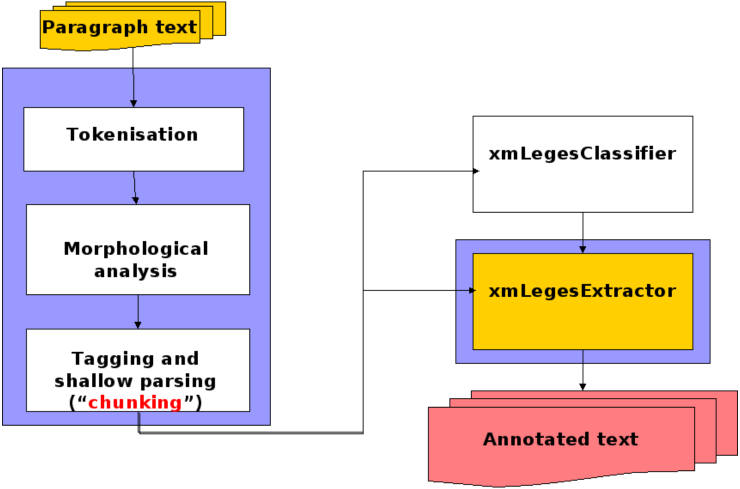


Experimental Results

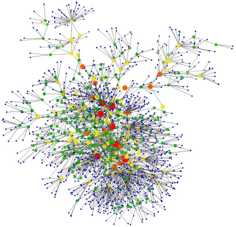
Data set composed by 473 legal text paragraphs

Provision Class	Success	Partial Success	Failure
Repeal	95.71%	2.86%	1.43%
Prohibition	73.33%	26.67%	–
Insertion	97.48%	1.68%	0.84%
Duty	88.89%	11.11%	–
Permission	66.67%	20%	13.33%
Penalty	47.93%	45.45%	6.61%
Substitution	96.40%	3.60%	–
Tot.	**82.09%**	**15.35%**	**2.56%**

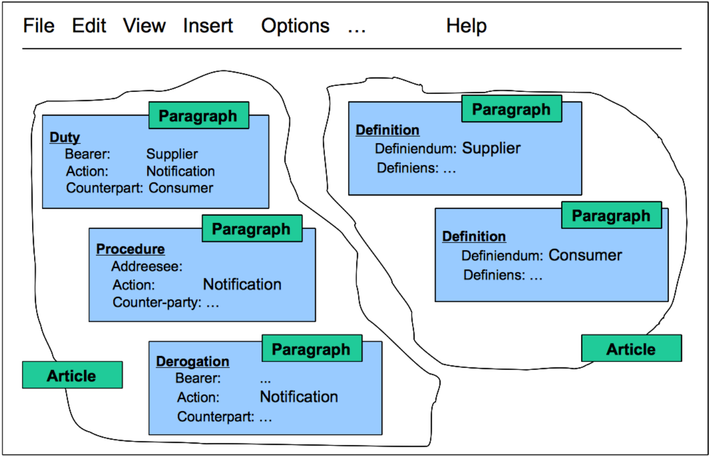
System FlowChart



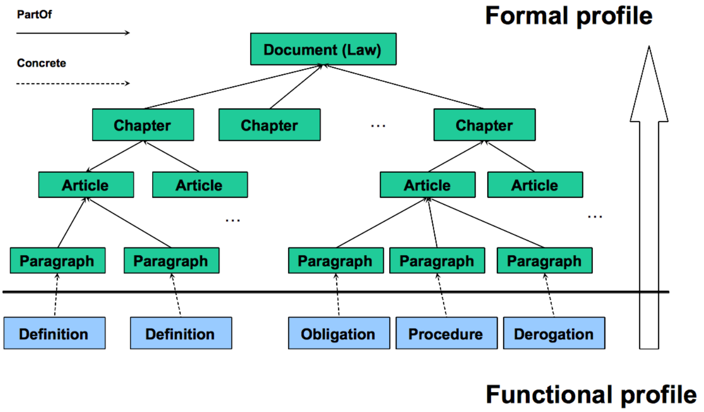
Semantic annotation
Top-Down Approach



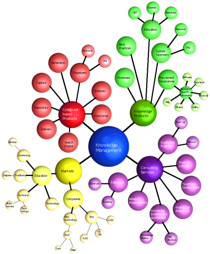
Visual semantic environment for drafting a new bill



[Biagioli et al., 2007]


PartOf
Concrete
Formal profile
Document (Law)
Chapter
Chapter
...
Chapter
Article
Article
Article
Article
...
...
Paragraph
Paragraph
Paragraph
Paragraph
Paragraph
Definition
Definition
Obligation
Procedure
Derogation
Functional profile

The Linked Data approach to the Semantic Web recommends to include Links between resources



Different vocabularies to represent the same type of entity

Mapping between Knowledge Resources (Thesauri/Ontology concepts)

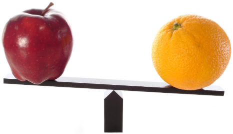
Interoperability

Thesaurus Mapping ($\mathcal{TM}$)

Definition

The process of identifying terms, concepts and hierarchical relationships that are approximately equivalent between thesauri

How to define and measure equivalence between concepts?



Concepts equivalence

Definition (*Instance-based equivalence*)

Two concepts are deemed to be equivalent if they are associated with, or classify the same set of objects

Definition (*Schema-based equivalence*)

Two concepts are deemed to be equivalent if there exists a similarity among their features

Concepts equivalence

Definition (*Schema-based equivalence*)

Two concepts are deemed to be equivalent if there exists a similarity among their features

Our proposal for Thesaurus Mapping formal characterization

Thesaurus Mapping ($\mathcal{TM}$) characterized as a problem of Information Retrieval ($\mathcal{IR}$)

- $\mathcal{IR}$: retrieve documents, in a document collection, better matching the semantics of a query
- $\mathcal{TM}$: retrieve concepts, in a target thesaurus, better matching the semantics of a given concept in a source thesaurus

$\mathcal{TM}$		$\mathcal{IR}$
Concept in source thesaurus	$\Longleftrightarrow$	Query
Concept in target thesaurus	$\Longleftrightarrow$	Document

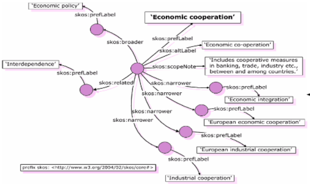
Isomorphism between $\mathcal{TM}$ and $\mathcal{IR}$

$$\mathcal{TM} \Longleftrightarrow \mathcal{IR}$$

$\mathcal{TM}$ formal characterization

Source thesaurus concepts

$Q = \{ q_i \} \qquad q_i = [x_1, x_2, \dots x_n] \in R^n$



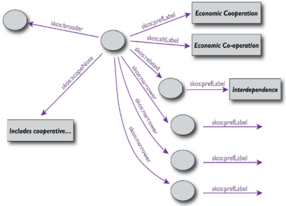
$\mathcal{TM} = [Q, D, F, R(q_i, d_j)]$

Framework

$F \equiv (R^n, dist/sim)$

Ranking Function

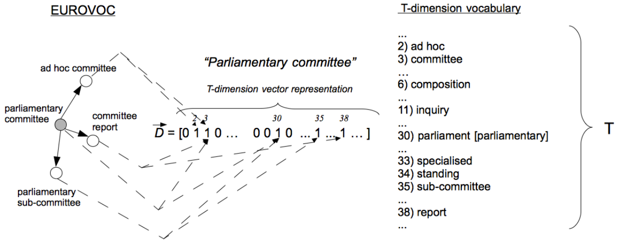
$dist/sim = R(q_i, d_j) \geq 0$



Target thesaurus concepts

$D = \{ d_j \} \qquad d_j = [y_1, y_2, \dots y_n] \in R^n$

Logical Views of concepts in source (Q) and target (D) thesauri



Pre-processing

- word stemming
- stopwords elimination

Vector $\vec{d} = [x_1, \ldots, x_{|T|}], \quad x_i \in \{0, 1\}$ composed by

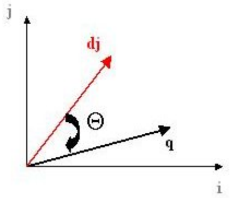
- the term itself
- relevant terms in its definition and in the alternative labels
- related thesaural concept terms

- T : dimension of the target thesaurus vocabulary
- x_i : presence/absence of the i^{th} vocabulary term in the concept $\vec{d}$.

The proposed Ranking Function (R)

Thesaural concepts similarity is measured as correlation between the related vectors

$$R = sim(\vec{q}, \vec{d}) = \frac{\vec{q} \times \vec{d}}{|\vec{q}| \cdot |\vec{d}|}$$



$|\vec{q}|$ and $|\vec{d}|$ are the norms of the vectors representing concepts in source and target thesauri.

A machine learning technique for conceptual mapping prediction

Criterion to predict matching concepts over a similarity measure

- Heuristic thresholds over $sim(q_i, d_j)$:

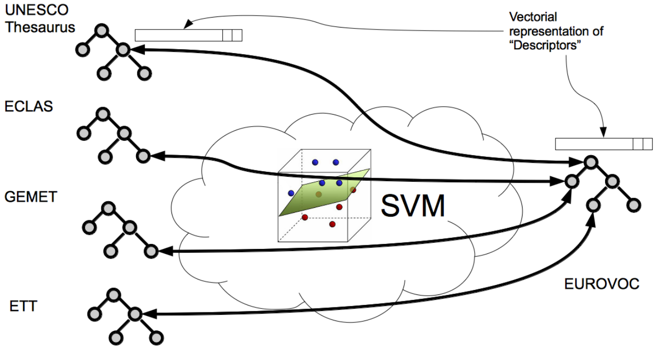
if	$sim(q_i, d_j) < T_1$	$\Rightarrow$	No Match
if	$T_1 < sim(q_i, d_j) < T_2$	$\Rightarrow$	partial match (broad or narrowMatch)
if	$T_2 < sim(q_i, d_j)$	$\Rightarrow$	exactMatch

  Problems in generalization capabilities out of the matching examples used to tune the heuristics.

- Generalization capabilities is introduced by a ML technique

SVM for conceptual mapping prediction

Support Vector Machine (SVM) trained to classify a descriptors pair as {Match (+1), no-Match (-1)}.



Training set for conceptual mapping prediction

Vectors Φ_i of features deemed representative for $(\vec{q}, \vec{d_i})$ conceptual mapping, including

- the similarity measure $sim(\vec{q}, \vec{d_i})$
- the logical view of the target descriptor $\vec{d_i}$
- a relevance judgment $y = \{+1(\text{Match}), -1(\text{NoMatch})\}$ for $\vec{d_i}$ on $\vec{q}$

$$\Phi_i = \langle\ \langle sim(\vec{d_i}, \vec{q}), \vec{d_i}\rangle\ ,\ y_i\rangle$$

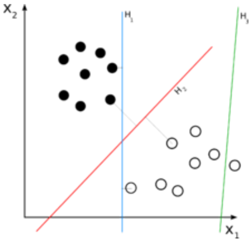
Training set for conceptual mapping prediction

Vectors Φ_i of features deemed representative for $(\vec{q}, \vec{d_i})$ conceptual mapping, including

- the similarity measure $sim(\vec{q}, \vec{d_i})$
- the logical view of the target descriptor $\vec{d_i}$
- a relevance judgment $y = \{+1(\text{Match}), -1(\text{NoMatch})\}$ for $\vec{d_i}$ on $\vec{q}$

$$\Phi_i = \langle \ \langle sim(\vec{d_i}, \vec{q}), \vec{d_i} \rangle \ , \ y_i \rangle$$

Distance of the examples wrt a separating surface gives a measure of prediction confidence



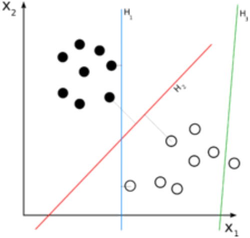
Training set for conceptual mapping prediction

Vectors Φ_i of features deemed representative for $(\vec{q}, \vec{d_i})$ conceptual mapping, including

- the similarity measure $sim(\vec{q}, \vec{d_i})$
- the logical view of the target descriptor $\vec{d_i}$
- a relevance judgment $y = \{+1(\text{Match}), -1(\text{NoMatch})\}$ for $\vec{d_i}$ on $\vec{q}$

$$\Phi_i = \langle \ \langle sim(\vec{d_i}, \vec{q}), \vec{d_i} \rangle \ , \ y_i \rangle$$

Distance of the examples wrt a separating surface gives a measure of prediction confidence

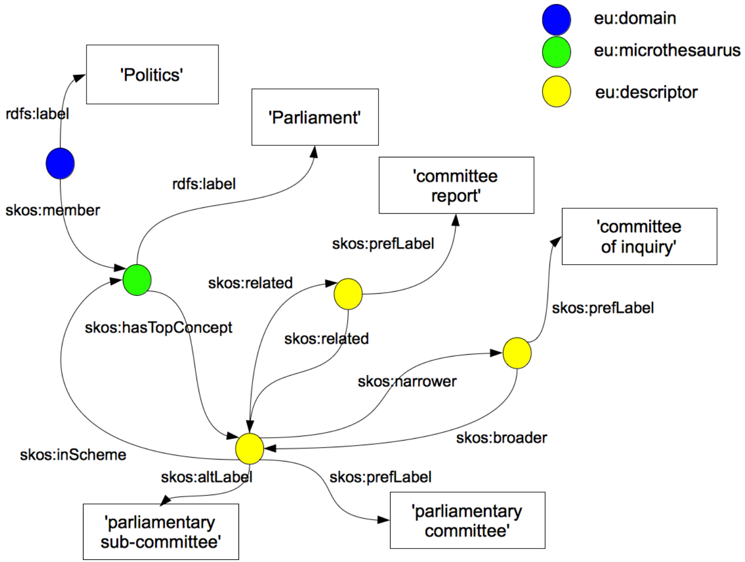


The best ranked concept is chosen as the predicted matching concept

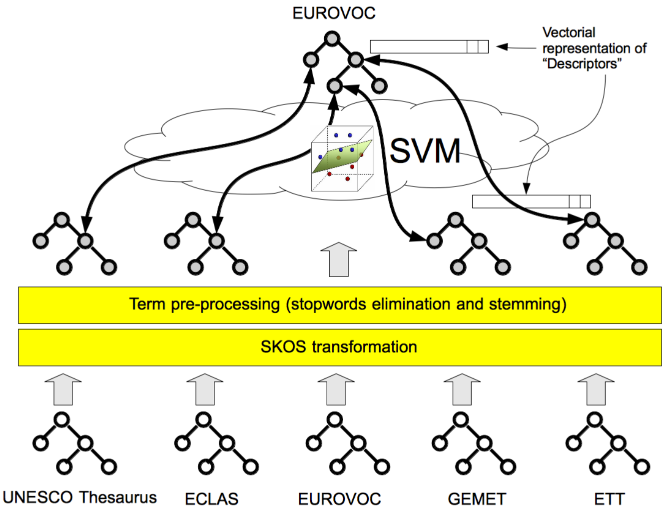
Interoperability among Thesauri: the case study

- EUROVOC the main EU thesaurus considering issues of specific and common interest for the EU and its Member States
- ECLAS the European Commission Central Libraries thesaurus
- GEMET GEneral Multilingual Environmental Thesaurus
- UNESCO Thesaurus developed by the United Nations Educational, Scientific and Cultural Organisation
- European Training Thesaurus (ETT) a thesaurus providing support to indexing and retrieval vocational education and training documentation in the European Union

Excerpt of Eurovoc SKOS representation



Workflow



EUROVOC
Vectorial representation of "Descriptors"
SVM
Term pre-processing (stopwords elimination and stemming)
SKOS transformation
UNESCO Thesaurus
ECLAS
EUROVOC
GEMET
ETT

The "Gold Standard" data set

Thesauri	`skos:exactMatch` **relations**
EUROVOC-ETT	131
EUROVOC-UNESCO	93
EUROVOC-ECLAS	143
EUROVOC-GEMET	28
Total exact match	395

Experimental Results

altLabel	Related concepts	*Accuracy*
no	no	83,87%
yes	no	93,55%
no	yes	100%
yes	yes	100%

EUROVOC-UNESCO mapping

altLabel	Related concepts	*Accuracy*
no	no	87,02%
yes	no	95,42%
no	yes	100%
yes	yes	100%

EUROVOC-ETT mapping

altLabel	Related concepts	*Accuracy*
no	no	93,00%
yes	no	93,71%

EUROVOC-ECLAS mapping

altLabel	Related concepts	Accuracy
no	no	100,00%

EUROVOC-GEMET mapping

Conclusions

Semantic annotation of legal texts using AI approaches

Bottom-up semantic annotation

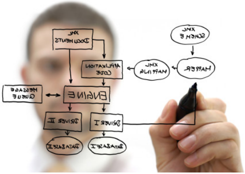
- Machine learning (SVM)
- NLP (Chunking)

Top-down semantic annotation

- Model-driven legal drafting

Interoperability between Knowledge Models and between Data

- Machine learning to establish semantic similarity between concepts



Thanks for your attention!

enrico.francesconi@ittig.cnr.it
enrico.francesconi@publications.europa.eu

Biagioli, C., Cappelli, A., Francesconi, E., and Turchi, F. (2007).
Law making environment: perspectives.
In *Proceedings of the V Legislative XML Workshop*, pages 267–281. European Press Academic Publishing.

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., and Soria, C. (2005).
Automatic semantics extraction in law documents.
In *International Conference on Artificial Intelligence and Law*, pages 133–139.

de Maat, E., Krabben, K., and Winkels, R. (2010).
Machine learning versus knowledge based classification of legal texts.
In *Proceedings of the Jurix Conference: Legal Knowledge and Information Systems*, pages 87–96, The Netherlands. IOS Press.

Francesconi, E. and Passerini, A. (2007).

Automatic classification of provisions in legislative texts.
International Journal on Artificial Intelligence and Law, 15(1):1–17.