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

- prv:TemporalAmendment
- prv:ExtensionAmendment
- prv:ContentAmendment
- prv:Constitutive
- prv:Regulative
- prv:Rule
- prv:RuleOnRule
- prv:Provision
Regulatives provisions

- prv:Permission
- prv:Prohibition
- prv:Duty
- prv:Right
- prv:Violation
- prv:Redress

prv:RuleOnAction

prv:Regulative

prv:Remedy
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 [...]

[...]
<table>
<thead>
<tr>
<th><strong>Art. 5</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> 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 [...]</td>
</tr>
<tr>
<td><strong>2.</strong> 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 [...]</td>
</tr>
<tr>
<td><strong>3.</strong> At any time during the contractual relationship the consumer is entitled, at his request, to receive the contractual terms and conditions on paper. [...]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Art. 6</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> 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 [...]</td>
</tr>
</tbody>
</table>
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)

[...]
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
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 occurring in many different documents, being less discriminative);

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
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 \(\implies H(D_w) = 0\)

where \(H(D) = \sum_{i=1}^{|C|} -p_i \log_2(p_i)\)
Data set of 582 examples (fragments of text containing a provision), belonging to 11 classes

<table>
<thead>
<tr>
<th>Class labels</th>
<th>Provision Types</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_0$</td>
<td>Repeal</td>
<td>70</td>
</tr>
<tr>
<td>$c_1$</td>
<td>Definition</td>
<td>10</td>
</tr>
<tr>
<td>$c_2$</td>
<td>Delegation</td>
<td>39</td>
</tr>
<tr>
<td>$c_3$</td>
<td>Delegification</td>
<td>4</td>
</tr>
<tr>
<td>$c_4$</td>
<td>Duty</td>
<td>13</td>
</tr>
<tr>
<td>$c_5$</td>
<td>Exception</td>
<td>18</td>
</tr>
<tr>
<td>$c_6$</td>
<td>Inserting</td>
<td>121</td>
</tr>
<tr>
<td>$c_7$</td>
<td>Prohibition</td>
<td>59</td>
</tr>
<tr>
<td>$c_8$</td>
<td>Permission</td>
<td>15</td>
</tr>
<tr>
<td>$c_9$</td>
<td>Penalty</td>
<td>122</td>
</tr>
<tr>
<td>$c_{10}$</td>
<td>Substitution</td>
<td>111</td>
</tr>
</tbody>
</table>
Naïve Bayes

Using paragraphs full text

<table>
<thead>
<tr>
<th>N terms with max InfoGain</th>
<th>Train Accuracy</th>
<th>LOO Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>90.7%</td>
<td>86.9%</td>
</tr>
<tr>
<td>50</td>
<td>89.3%</td>
<td>86.9%</td>
</tr>
</tbody>
</table>

Excluding quoted text ("misleading text")

<table>
<thead>
<tr>
<th>N terms with max InfoGain</th>
<th>Train Accuracy</th>
<th>LOO Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>95.5%</td>
<td>88.6%</td>
</tr>
<tr>
<td>250</td>
<td>94.3%</td>
<td>88.1%</td>
</tr>
</tbody>
</table>
Using paragraphs **full text**

<table>
<thead>
<tr>
<th>Train Accuracy</th>
<th>LOO Accuracy</th>
<th>N terms with max InfoGain</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>91.2%</td>
<td>1000</td>
</tr>
<tr>
<td>100%</td>
<td>91.9%</td>
<td>500</td>
</tr>
</tbody>
</table>

Excluding quoted text ("misleading text")

<table>
<thead>
<tr>
<th>Train Accuracy</th>
<th>LOO Accuracy</th>
<th>N terms with max InfoGain</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.8%</td>
<td>92.1%</td>
<td>all</td>
</tr>
<tr>
<td>99.8%</td>
<td>92.1%</td>
<td>1000</td>
</tr>
</tbody>
</table>
Chunking and SVM

Text representation using **linguistic structures of higher level of abstraction**

Using paragraphs **full text**

<table>
<thead>
<tr>
<th>Train Accuracy</th>
<th>LOO Accuracy</th>
<th>N terms with max InfoGain</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.7%</td>
<td>92.4%</td>
<td>all</td>
</tr>
<tr>
<td>99.7%</td>
<td>92.4%</td>
<td>100</td>
</tr>
</tbody>
</table>

Excluding quoted text (**“misleading text”**)  

<table>
<thead>
<tr>
<th>Train Accuracy</th>
<th>LOO Accuracy</th>
<th>N terms with max InfoGain</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.7%</td>
<td>92.7%</td>
<td>all</td>
</tr>
<tr>
<td>99.7%</td>
<td>92.7%</td>
<td>500</td>
</tr>
</tbody>
</table>
Comparison of the Results

- SVM con chunker
- SVM senza chunker
- Naive Bayes senza chunker

LOO Accuracy
1. A controller intending to process personal data falling within the scope of application of this Act shall have to notify the Garante thereof…

**Provision type:** “Duty”

**Attributes:**
- **Bearer:** “Controller”
- **Action:** “Notification”
- **Counterpart:** “Garante”
- **Object:** “Process personal data”
## Experimental Results

Data set composed by 473 legal text paragraphs

<table>
<thead>
<tr>
<th>Provision Class</th>
<th>Success</th>
<th>Partial Success</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeal</td>
<td>95.71%</td>
<td>2.86%</td>
<td>1.43%</td>
</tr>
<tr>
<td>Prohibition</td>
<td>73.33%</td>
<td>26.67%</td>
<td>–</td>
</tr>
<tr>
<td>Insertion</td>
<td>97.48%</td>
<td>1.68%</td>
<td>0.84%</td>
</tr>
<tr>
<td>Duty</td>
<td>88.89%</td>
<td>11.11%</td>
<td>–</td>
</tr>
<tr>
<td>Permission</td>
<td>66.67%</td>
<td>20%</td>
<td>13.33%</td>
</tr>
<tr>
<td>Penalty</td>
<td>47.93%</td>
<td>45.45%</td>
<td>6.61%</td>
</tr>
<tr>
<td>Substitution</td>
<td>96.40%</td>
<td>3.60%</td>
<td>–</td>
</tr>
<tr>
<td><strong>Tot.</strong></td>
<td><strong>82.09%</strong></td>
<td><strong>15.35%</strong></td>
<td><strong>2.56%</strong></td>
</tr>
</tbody>
</table>
System FlowChart

Paragraph text

Tokenisation

Morphological analysis

Tagging and shallow parsing ("chunking")

xmlLegesClassifier

xmlLegesExtractor

Annotated text
Semantic annotation
Top-Down Approach
Visual semantic environment for drafting a new bill

[Biagioli et al., 2007]
Model Driven Legislative Drafting

Enrico Francesconi
AI and Law-Semantic Annotation of Legal Texts
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 ($\mathcal{T}M$)

**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

<table>
<thead>
<tr>
<th>$\mathcal{TM}$</th>
<th>$\mathcal{IR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept in source thesaurus</td>
<td>Query</td>
</tr>
<tr>
<td>Concept in target thesaurus</td>
<td>Document</td>
</tr>
</tbody>
</table>
Isomorphism between $TM$ and $IR$
$TM$ formal characterization

**Source thesaurus concepts**

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

$TM = [Q, D, F, R(q_i, d_j)]$

**Framework**

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

**Ranking Function**

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

**Target thesaurus concepts**

$D = \{ d_j \}$  
$d_j = [y_1, y_2, \ldots, y_n] \in R^n$
Logical Views of concepts in source ($Q$) and target ($D$) thesauri

Pre-processing
- word stemming
- stopwords elimination

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

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

$|\vec{q}|$ 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 $\text{sim}(q_i, d_j)$:
  - if $\text{sim}(q_i, d_j) < T_1 \Rightarrow$ No Match
  - if $T_1 < \text{sim}(q_i, d_j) < T_2 \Rightarrow$ partial match (broad or narrowMatch)
  - if $T_2 < \text{sim}(q_i, d_j) \Rightarrow$ exactMatch

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

- Generalization capabilities is introduced by a ML technique
Support Vector Machine (SVM) trained to classify a descriptors pair as \{Match (+1), no-Match (-1)\}. 
Training set for conceptual mapping prediction

Vectors $\Phi_i$ of features deemed representative for $(\vec{q}, \vec{d}_i)$ conceptual mapping, including

- the similarity measure $\text{sim}(\vec{q}, \vec{d}_i)$
- 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 $\vec{q}$

$$\Phi_i = \langle \langle \text{sim}(\vec{d}_i, \vec{q}), \vec{d}_i \rangle, y_i \rangle$$
Training set for conceptual mapping prediction

Vectors $\Phi_i$ of features deemed representative for $(\vec{q}, \vec{d}_i)$ conceptual mapping, including

- the similarity measure $\text{sim}(\vec{q}, \vec{d}_i)$
- 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 $\vec{q}$

$$\Phi_i = \langle \langle \text{sim}(\vec{d}_i, \vec{q}), \vec{d}_i \rangle, y_i \rangle$$

Distance of the examples wrt a separating surface gives a measure of prediction confidence.
Vectors $\Phi_i$ 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

Enrico Francesconi  AI and Law-Semantic Annotation of Legal Texts
Workflow

Term pre-processing (stopwords elimination and stemming)

SKOS transformation

UNESCO Thesaurus  ECLAS  EUROVOC  GEMET  ETT
The “Gold Standard” data set

<table>
<thead>
<tr>
<th>Thesauri</th>
<th>skos:exactMatch relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUROVOC-ETT</td>
<td>131</td>
</tr>
<tr>
<td>EUROVOC-UNESCO</td>
<td>93</td>
</tr>
<tr>
<td>EUROVOC-ECLAS</td>
<td>143</td>
</tr>
<tr>
<td>EUROVOC-GEMET</td>
<td>28</td>
</tr>
<tr>
<td><strong>Total exact match</strong></td>
<td><strong>395</strong></td>
</tr>
</tbody>
</table>
### Experimental Results

<table>
<thead>
<tr>
<th>altLabel</th>
<th>Related concepts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>83.87%</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>93.55%</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>100%</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>100%</td>
</tr>
</tbody>
</table>

**EUROVOC-UNESCO mapping**

<table>
<thead>
<tr>
<th>altLabel</th>
<th>Related concepts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>87.02%</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>95.42%</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>100%</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>altLabel</th>
<th>Related concepts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>93.00%</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>93.71%</td>
</tr>
</tbody>
</table>

**EUROVOC-ETT mapping**

<table>
<thead>
<tr>
<th>altLabel</th>
<th>Related concepts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>yes</td>
<td>100%</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>altLabel</th>
<th>Related concepts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

**EUROVOC-ECLAS mapping**

<table>
<thead>
<tr>
<th>altLabel</th>
<th>Related concepts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>yes</td>
<td>100%</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>100%</td>
</tr>
</tbody>
</table>

**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
Law making environment: perspectives.

Automatic semantics extraction in law documents.
In *International Conference on Artificial Intelligence and Law*, pages 133–139.

Machine learning versus knowledge based classification of legal texts.


Enrico Francesconi
AI and Law-Semantic Annotation of Legal Texts