NATURAL LANGUAGE PROCESSING AND
LEGAL KNOWLEDGE EXTRACTION

Simonetta Montemagni, Giulia Venturi
Istituto di Linguistica Computazionale «Antonio Zampolli»
(ILC-CNR)
Bridging the gap between text and knowledge: the crucial role of NLP tools

Knowledge is mostly conveyed through text
- Content access requires understanding the linguistic structure

We need a bridge to overcome the gap between text and knowledge

Technologies based on Natural Language Processing allows
- accessing the domain-specific knowledge contained in texts
- structuring the textual content
From text to knowledge: the main challenge in the legal domain

One of the main obstacles to progress in the field of artificial intelligence and law is the **natural language barrier**

L. Thorne McCarty, International Conference on AI and Law (ICAIL-2007)

- Raw materials of the law are embodied in natural language (cases, statutes, regulations, etc.)
- Legal knowledge is heavily intertwined with natural language and common sense and therefore inherits all the hard problems that these imply
- Knowledge-based legal information systems need to access the content embedded in legal texts
- Legal search engine
- gathering Italian different sources of law (case laws, legislation, jurisprudence, journals, etc.)
IUSEXPLORER: an example of word search query

Ambiguity between the verb and the noun
IUSEXPLORER: an example of word search query

It returns the **single word** (damage and patrimonial), the **multi-word** and also the **negation**

danno patrimoniale
(patrimonial damage)
Advanced search engine which provides customers with access to billions of searchable documents

It is still linguistically rudimentary
- it does not exploit the potential offered by language technologies
- it does not support semantic queries allowing an advanced access to documents

Need for increasingly sophisticated applications based on Natural Language Processing technologies for effectively accessing the content embedded in texts
Summary

- From text to knowledge
  - The general approach

- Natural Language Processing tools
  - What and what for

- The main challenges of the legal domain

- Legal Knowledge Extraction
  - Identification and extraction of domain-relevant knowledge
  - Semantic annotation of legal texts
From text to knowledge: the general approach

Textual content
(implicit knowledge)

Structured knowledge
(explicit knowledge)

Dynamic content structuring

Linguistic annotation

Knowledge extraction
Natural Language Processing techniques and knowledge extraction

Linguistic annotation
- text
- Sentence Splitter
- Tokenizer
- Morphological analyzer
- Part-Of-Speech Tagger
- Dependency parser

Knowledge extraction tools

Structured knowledge

Lexico-semantic resources

Balanced Cooperative Approach
Linguistic annotation tools: what

- Linguistic annotation
  - the process in charge of reconstructing and making explicit the linguistic structure underlying texts

- State-of-the-art tools are based on machine-learning algorithms
  - Annotation process as probabilistic classification task

- Basic requirements
  - robustness to minimize failures due to lexical gaps, particularly complex linguistic constructions as well as ill-formed input
  - accuracy of achieved results
  - efficiency to deal with huge amounts of textual data
  - portability to different domains, textual genres, linguistic registers, other languages
  - incrementality of analysis
Linguistic annotation: an incremental process

- **Sentence Splitter**: Splits the text into sentences
- **Tokenizer**: Segments each sentence into orthographic units (tokens)
- **Morphological analyzer**: Assigns the possible morphological analyses to each token
- **PoS Tagger**: Selects the appropriate morphological interpretation in the specific context
- **Dependency parser**: Identifies dependency relations between tokens (e.g. subject, object, etc.)
Linguistic annotation: an example

Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)
Linguistic annotation: an example

- Il danno non poteva essere sottovalutato. (The damage could not be underestimated.)
- Il sig. Rossi decise perciò di chiamare l’avvocato. (Mr. Rossi decided therefore to call the lawyer.)
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be underestimated.)
- Il sig. Rossi decise perciò di chiamare l’avvocato. (Mr. Rossi decided therefore to call the lawyer.)

<table>
<thead>
<tr>
<th>id</th>
<th>form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Il</td>
</tr>
<tr>
<td>2</td>
<td>danno</td>
</tr>
<tr>
<td>3</td>
<td>non</td>
</tr>
<tr>
<td>4</td>
<td>poteva</td>
</tr>
<tr>
<td>5</td>
<td>essere</td>
</tr>
<tr>
<td>6</td>
<td>sottovalutato</td>
</tr>
</tbody>
</table>

"CoNLL" tabular representation schema
Linguistic annotation: an example

Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be underestimated.)
- Il sig. Rossi decise perciò di chiamare l’avvocato. (Mr. Rossi decided therefore to call the lawyer.)

<table>
<thead>
<tr>
<th>id</th>
<th>form</th>
<th>lemma</th>
<th>PoS</th>
<th>Feats</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>il</td>
<td>il</td>
<td>RD</td>
<td>MS</td>
</tr>
<tr>
<td>2</td>
<td>danno</td>
<td>danno;dare</td>
<td>S;V</td>
<td>MS;P3IP</td>
</tr>
<tr>
<td>3</td>
<td>non</td>
<td>non</td>
<td>BN</td>
<td>NULL</td>
</tr>
<tr>
<td>4</td>
<td>poteva</td>
<td>potere</td>
<td>V</td>
<td>S3II</td>
</tr>
<tr>
<td>5</td>
<td>essere</td>
<td>essere</td>
<td>V</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>sottovalutato</td>
<td>sottovalutare</td>
<td>V</td>
<td>MSPR</td>
</tr>
</tbody>
</table>

"CoNLL" tabular representation schema
Linguistic annotation: an example

Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l’avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be underestimated.)
- Il sig. Rossi decise perciò di chiamare l’avvocato. (Mr. Rossi decided therefore to call the lawyer.)

<table>
<thead>
<tr>
<th>id</th>
<th>form</th>
<th>lemma</th>
<th>PoS</th>
<th>Feats</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>il</td>
<td>il</td>
<td>RD</td>
<td>MS</td>
</tr>
<tr>
<td>2</td>
<td>danno</td>
<td>danno</td>
<td>S</td>
<td>MS</td>
</tr>
<tr>
<td>3</td>
<td>non</td>
<td>non</td>
<td>BN</td>
<td>NULL</td>
</tr>
<tr>
<td>4</td>
<td>poteva</td>
<td>potere</td>
<td>V</td>
<td>S3II</td>
</tr>
<tr>
<td>5</td>
<td>essere</td>
<td>essere</td>
<td>V</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>sottovalutato</td>
<td>sottovalutare</td>
<td>V</td>
<td>MSPR</td>
</tr>
</tbody>
</table>

”CoNLL” tabular representation schema
Linguistic annotation: an example

Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be underestimated.)
- Il sig. Rossi decise perciò di chiamare l'avvocato. (Mr. Rossi decided therefore to call the lawyer.)

<table>
<thead>
<tr>
<th>id</th>
<th>form</th>
<th>lemma</th>
<th>PoS</th>
<th>Feats</th>
<th>head</th>
<th>DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Il</td>
<td>il</td>
<td>RD</td>
<td>MS</td>
<td>2</td>
<td>DET</td>
</tr>
<tr>
<td>2</td>
<td>danno</td>
<td>danno</td>
<td>S</td>
<td>MS</td>
<td>6</td>
<td>SUBJ_PASS</td>
</tr>
<tr>
<td>3</td>
<td>non</td>
<td>non</td>
<td>BN</td>
<td>NULL</td>
<td>6</td>
<td>NEG</td>
</tr>
<tr>
<td>4</td>
<td>poteva</td>
<td>potere</td>
<td>V</td>
<td>S3II</td>
<td>6</td>
<td>MODAL</td>
</tr>
<tr>
<td>5</td>
<td>essere</td>
<td>essere</td>
<td>V</td>
<td>F</td>
<td>6</td>
<td>AUX</td>
</tr>
<tr>
<td>6</td>
<td>sottovalutato</td>
<td>sottovalutare</td>
<td>V</td>
<td>MSPR</td>
<td>0</td>
<td>ROOT</td>
</tr>
</tbody>
</table>

"CoNLL" tabular representation schema
Linguistic annotation: an example

Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l’avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be underestimated.)
- Il sig. Rossi decise perciò di chiamare l’avvocato. (Mr. Rossi decided therefore to call the lawyer.)

<table>
<thead>
<tr>
<th>id</th>
<th>form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Il</td>
</tr>
<tr>
<td>2</td>
<td>danno</td>
</tr>
<tr>
<td>3</td>
<td>non</td>
</tr>
<tr>
<td>4</td>
<td>poteva</td>
</tr>
<tr>
<td>5</td>
<td>essere</td>
</tr>
<tr>
<td>6</td>
<td>sottovalutato</td>
</tr>
</tbody>
</table>

- **CoNLL** tabular representation schema

```
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l’avvocato. (The damage could not be underestimated. Mr. Rossi decided therefore to call the lawyer.)
```
Linguistic annotation: what for

Linguistic annotation plays a crucial role in accessing the content of texts by making it explicit the linguistic structure through which knowledge is encoded.

Starting point for several Knowledge Extraction tasks:
- Extracting domain-relevant knowledge
- Structuring the extracted knowledge in semantic resources, e.g., lexicons, thesauri, domain-specific ontologies (Ontology Learning)
- Semantic indexing of text collections on the basis of the extracted knowledge

Linguistic annotation and knowledge extraction:
- Increasingly complex knowledge extraction tasks differentially exploit individual levels of linguistic annotation.
From text to knowledge: the general approach

Textual content (implicit knowledge)

Dynamic content structuring

Knowledge extraction

Structured knowledge (explicit knowledge)

Linguistic annotation
The legal domain: the main challenges

- The typical **knowledge acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure

- The **peculiarity of legal language and its impact on NLP tools**
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) with respect to ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?

- **Discriminate between legal and regulated domain knowledge**
  - By its very nature, law deals with behaviour in the world: domain independent concepts of law are tainted with concepts referring to the world the legal domain is about
The knowledge acquisition bottleneck

- Technologies in the area of knowledge management are typically confronted with the problem of processing linguistic structure
  - Particularly relevant in the legal domain, where law is strictly dependent on its linguistic expression

- Why legal language processing?
  - “Why parse statutes? To extract their logical structure, to refine the semantics of the domain, to develop a domain ontology” (McCarty, 2009)

- What are the domain-specific issues to be addressed when processing legal language?
  - Whether and to what extent legal language differs from ordinary language
  - Impact of recorded differences on the performance of NLP tools
The legal domain: the main challenges

- The typical knowledge acquisition bottleneck
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure

- The peculiarity of legal language and its impact on NLP tools
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) with respect to ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?

- Discriminate between legal and regulated domain knowledge
  - By its very nature, law deals with behaviour in the world: domain independent concepts of law are tainted with concepts referring to the world the legal domain is about
The peculiarity of legal language

- Legal texts differ significantly with respect to ordinary language texts
  - typically correlated with syntactic complexity

- Differences recorded at different annotation levels
  - long sentences wrt newswire texts
The peculiarity of legal language

- Legal texts differ significantly with respect to ordinary language texts
  - typically correlated with syntactic complexity
- Differences recorded at different annotation levels
  - long sentences wrt newswire texts

**Italian:**
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

**English:**
- a corpus of newspapers
- a collection of laws enacted by the European Commission

![Graph showing average sentence length for Italian and English legal texts](image)
The peculiarity of legal language

- Legal texts differ significantly with ordinary language texts
  - typically correlated with syntactic complexity

- Differences recorded at different annotation levels
  - long sentences wrt newswire texts
  - high % of prepositions and low % of verbs, adverbs, pronouns

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

<table>
<thead>
<tr>
<th></th>
<th>IT_newspapers</th>
<th>IT_regional&amp;national_laws</th>
<th>IT_EU_laws</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepositions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nouns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverbs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronouns</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

<table>
<thead>
<tr>
<th></th>
<th>EN_newspapers</th>
<th>EN_EU_laws</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepositions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nouns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverbs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronouns</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The peculiarity of legal language

- Legal texts differ significantly with respect to ordinary texts
  - typically correlated with syntactic complexity

- Differences recorded at different annotation levels
  - long sentences wrt newswire texts
  - high % of prepositions and low % of verbs, adverbs, pronouns
  - deep sequences of embedded prepositional complement chains

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

![Bar graph comparison of average depth of embedded complement 'chains' for IT and EN corpora.](chart)
The peculiarity of legal language

- Legal texts differ significantly with respect to ordinary language texts
  - typically correlated with syntactic complexity

- Differences recorded at different annotation levels
  - long sentences wrt newswire texts
  - high % of prepositions and low % of verbs, adverbs, pronouns
  - deep sequences of embedded prepositional complement chains

**Italian:**
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

**English:**
- a corpus of newspapers
- a collection of laws enacted by the European Commission

---

Il Consiglio è giunto ad un accordo sui contributi dei singoli Stati membri all’adempimento dell’impegno globale di riduzione delle emissioni della Comunità nelle conclusioni del Consiglio del 16 giugno 1998. (The Council agreed upon the contributions of each Member State to the overall Community reduction commitment in the Council conclusions of 16 June 1998)
The peculiarity of legal language

- Legal texts differ significantly with ordinary language texts
  - typically correlated with syntactic complexity

- Differences recorded at different levels:
  - long sentences wrt newswire texts
  - high % of prepositions and low % of verbs, adverbs, pronouns
  - long sequences of consecutive prepositional complements
  - long dependency links
  - deep syntactic trees

- Statistical parsers have a drop in accuracy when analyzing long distance dependencies (McDonald and Nivre, 2007)
- Parse tree depth is a well-known feature reflecting sentence complexity
The impact of legal language on NLP tools

- What is the performance of state-of-the-art NLP tools on legal texts?
  - A key issue for all NLP-based Knowledge Extraction tasks
  - Generally speaking, a dramatic drop of accuracy is reported when syntactic parsers are tested on domains outside of the data from which they are trained or developed on

- Recently, two initiatives focused on dependency parsing of legal texts which represents a prerequisite for any advanced legal text processing task
  - Domain Adaptation Track at Evalita 2011 – Italian
  - SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts – Italian and English
  - both aimed at
    - obtaining a clear idea of the current performance of state-of-the-art dependency parsing systems against legal texts
    - investigating techniques for adapting state-of-the-art dependency parsing systems to the legal domain
The impact of legal language on NLP tools

- Results of the Dependency Parsing subtask of the SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts
  - Goal: testing the performance of general parsing systems on legal texts
- Accuracy results for Italian:

<table>
<thead>
<tr>
<th>Participant System</th>
<th>Newspaper test</th>
<th>Reg/Nat legal test</th>
<th>EU legal test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82.36</td>
<td>75.88</td>
<td>83.08</td>
</tr>
<tr>
<td>2</td>
<td>82.90</td>
<td>74.03</td>
<td>81.93</td>
</tr>
<tr>
<td>3</td>
<td>81.43</td>
<td>75.55</td>
<td>81.58</td>
</tr>
</tbody>
</table>

- Accuracy results for English:

<table>
<thead>
<tr>
<th>Participant System</th>
<th>Newspaper test</th>
<th>EU legal test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>88.81</td>
<td>78.90</td>
</tr>
</tbody>
</table>

For both Italian and English, lower performance of parsing systems on legal texts wrt newspapers

Different performances across different subvarietes of legal language
- Significant drops on the IT regional and national texts
- 2 out of the 3 participant systems do not show a significant drop of accuracy when tested on EU legal texts
The legal domain: the main challenges

- **The typical knowledge acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure

- **The peculiarity of legal language and its impact on NLP tools**
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) with respect to ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?

- **Discriminate between legal and regulated domain knowledge**
  - By its very nature, law deals with behaviour in the world: domain independent concepts of law are tainted with concepts referring to the world the legal domain is about
Discriminate between legal and regulated domain knowledge

«As any legal source – legislation, contracts, precedence-law – reveals immediately: the majority of concepts in an individual source refers to specific domains of social activities. These domains are called ‘world knowledge’.

Breuker & Hoekstra 2004

Domain-specific terms of law are tainted with terms referring to the world the legal domain is about

- e.g. national provision, fundamental principle & hazardous substance, active ingredient

Discriminating between legal and regulated domain terms and/or concepts is key in constructing a legal semantic resource

- It is closely related to the reusability and interoperability issue

«Therefore it is not surprise that one may find that many legal ontologies are mixtures of epistemological and ontological perspectives.

Breuker & Hoekstra 2004
According to the ontology design criteria, the level of generality in which concepts are organized is a distinctive characteristic.

Three different kinds of ontologies:
- Top or upper-level ontologies (general concepts)
- Core ontologies (top-level domain-specific concepts, e.g. legal)
- Domain-specific ontologies (which organize world knowledge)

Breuker & Hoekstra 2004: LRI-Core layers: foundational and legal core share ‘anchors’ (high level concepts typical for law)
From text to knowledge: the general approach

Textual content (implicit knowledge)

Structured knowledge (explicit knowledge)

Dynamic content structuring

Knowledge extraction

Linguistic annotation
Legal Knowledge Extraction: focus on …

- **Identification, extraction and structuring** of domain-relevant knowledge
  - Goal: constructing semantic resources such as domain-specific ontologies or lexicons

- **Semantic annotation** of legal texts
  - Goal: content-based access and querying
Legal Knowledge Extraction: focus on …

- **Identification, extraction and structuring** of domain-relevant knowledge
  - Goal: constructing semantic resources such as domain-specific ontologies or lexicons

Focus on the **Ontology Learning**

The construction of **Legal Ontologies** referred to as the «missing link» (Valente and Breuker, 2004) between Artificial Intelligence and Law and Legal Theory.

Key process since the emergence of the Semantic Web (Van Engers et al., 2008)
The various steps of Ontology Learning from texts can be arranged in a “layer cake” of increasingly complex subtasks

(Buitelaar, Cimiano and Magnini, 2005)

∀x, y (sufferFrom(x, y) → ill(x))
cure (dom:DOCTOR, range:DISEASE)
is_a (DOCTOR, PERSON)

DISEASE:=<Int,Ext,Lex>

{disease, illness}
disease, illness, hospital

Axioms & Rules
(Other) Relations
Taxonomy (Concept Hierarchies)
Concepts
Synonyms
Terms
Ontology Learning

First step of each Ontology Learning process:

Terminology Extraction

«Terms are linguistic realizations of domain-specific concepts and are therefore central to further, more complex tasks» (Buitelaar et al., 2005)

∀x, y (sufferFrom(x, y) → ill(x))
cure (dom:DOCTOR, range:DISEASE)
is_a (DOCTOR, PERSON)

DISEASE:=<Int,Ext,Lex>{disease, illness}
disease, illness, hospital

Axioms & Rules
(Other) Relations
Taxonomy (Concept Hierarchies)
Concepts
Synonyms
Terms
Ontology Learning: Terminology Extraction

- Terms may consist of
  - a single wordform so-called “simple” (or one-word) terms
    - e.g. artist
  - two or more wordforms, called “multi-word” (or complex) terms
    - e.g. art movement

- Term extraction process articulated into two fundamental steps:
  - identifying term candidates from text
  - filtering through the candidates to separate terms from non-terms

- Different statistical measures are used
  - For the extraction of simple terms: frequency occurrence distribution, measures of statistical relevance such as TF/IDF (Term Frequency/Inverse Term Frequency), etc.
  - For the extraction of multi-word terms: association strength measures such as Mutual Information, C-NC Value, Log-likelihood, etc.
The next step is the semantic structuring of the extracted terminology
- definition of concepts and relations between them

∀x, y (sufferFrom(x, y) → ill(x))
cure (dom:DOCTOR, range:DISEASE)
is_a (DOCTOR, PERSON)

DISEASE:=<Int,Ext,Lex>
{disease, illness}
disease, illness, hospital
The extracted terms are organized into fragments of taxonomical chains

- simple and multi-word terms are structured in a vertical hierarchy
- on the basis of their internal linguistic structure (head sharing)
Ontology Learning: to sum up

- Knowledge extraction in two steps:
  - **Term Extraction**: detection of single and multi-word terms
  - **Semantic Structuring**: definition of concepts and relations between them

![Diagram showing NLP tools and ontology learning process]
Ontology learning in the legal domain: so far …

- Overview of existing Legal Ontologies:

- Approaches to semi-automatically induce legal domain ontologies from texts
  - focus on definitions in German court decisions from which legal concepts are identified together with relevant terminology and relations
    - Walter and Pinkal (2006)
  - extraction of domain relevant terminology from which domain relevant concepts are derived together with relations linking them
    - Lame (2000, 2005): French
    - Saias and Quaresma (2005): Portuguese
    - Völker et al. (2008): Spanish
    - Lenci at al. (2009): Italian
  - ontology modelling
    - LKIF Core ontology (Hoekstra et al., 2007)
    - LOIS (Peters et al., 2005)
    - OPJK (Casellas, 2008)
    - DALOS (Agnoloni et al., 2009)
Ontology Learning: exemplifying Terminology Extraction

- Focus on the term extractor developed by ItaliaNLP Lab at ILC-CNR (Bonin et al., 2010)

- It follows a multilayered and contrastive approach to overcome the need to discriminate between legal and world knowledge
  - It singles out legal terms, e.g. *law, legislative decree* (legal knowledge), from regulated-domain terms, e.g. *consumer, hazardous substance* (world knowledge)

- Tested in different case studies
  - Corpus of environmental laws (Bonin et al., 2010)
    - EU Directives (394,088 tokens)
  - Case Law corpus (LIDER-Lab, Scuola Superiore Sant’Anna, Pisa)
    - Case law on personal offence (1,206,831 tokens)
  - Case Law corpus (Lazari & Venturi, 2012)
    - Case law on state liability (933,077 tokens)
The multi-layered architecture developed by the ItaliaNLP Lab
Linguistic annotation until the Part-Of-Speech and Lemmatization levels

E.g. Il piano nazionale di riduzione delle emissioni in nessun caso può esonerare un impianto dal rispetto della pertinente normativa comunitaria, compresa la direttiva 96/61/CE (The national emission reduction plan may under no circumstances exempt a plant from the provisions laid down in relevant Community legislation, including inter alia Directive 96/61/EC).

<table>
<thead>
<tr>
<th>Forma</th>
<th>Lemma</th>
<th>CPoSTag</th>
<th>PosTag</th>
<th>Tratti morfologici</th>
<th>Forma</th>
<th>Lemma</th>
<th>CPoSTag</th>
<th>PosTag</th>
<th>Tratti morfologici</th>
</tr>
</thead>
<tbody>
<tr>
<td>Il</td>
<td>il</td>
<td>R</td>
<td>RD</td>
<td>num=s</td>
<td>gen=m</td>
<td>un</td>
<td>un</td>
<td>R</td>
<td>RI</td>
</tr>
<tr>
<td>piano</td>
<td>piano</td>
<td>S</td>
<td>S</td>
<td>num=s</td>
<td>gen=m</td>
<td>impianto</td>
<td>impianto</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>nazionale</td>
<td>nazionale</td>
<td>A</td>
<td>A</td>
<td>num=s</td>
<td>gen=n</td>
<td>dal</td>
<td>da</td>
<td>E</td>
<td>EA</td>
</tr>
<tr>
<td>di</td>
<td>di</td>
<td>E</td>
<td>E</td>
<td>_</td>
<td>rispetto</td>
<td>rispetto</td>
<td>S</td>
<td>S</td>
<td>num=s</td>
</tr>
<tr>
<td>riduzione</td>
<td>riduzione</td>
<td>S</td>
<td>S</td>
<td>num=s</td>
<td>gen=f</td>
<td>della</td>
<td>di</td>
<td>E</td>
<td>EA</td>
</tr>
<tr>
<td>delle</td>
<td>di</td>
<td>E</td>
<td>EA</td>
<td>num=p</td>
<td>gen=f</td>
<td>pertinente</td>
<td>pertinente</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>emissioni</td>
<td>emissione</td>
<td>S</td>
<td>S</td>
<td>num=p</td>
<td>gen=f</td>
<td>normativa</td>
<td>normativa</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>in</td>
<td>in</td>
<td>E</td>
<td>E</td>
<td>_</td>
<td>comunitaria</td>
<td>comunitario</td>
<td>A</td>
<td>A</td>
<td>num=s</td>
</tr>
<tr>
<td>nessun</td>
<td>nessun</td>
<td>D</td>
<td>DI</td>
<td>num=s</td>
<td>gen=m</td>
<td>,</td>
<td>,</td>
<td>F</td>
<td>FF</td>
</tr>
<tr>
<td>caso</td>
<td>caso</td>
<td>S</td>
<td>S</td>
<td>num=s</td>
<td>gen=m</td>
<td>compresa</td>
<td>comprendere</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>può</td>
<td>potere</td>
<td>V</td>
<td>VM</td>
<td>num=s</td>
<td>per=3</td>
<td>mod=il</td>
<td>ten=p</td>
<td>la</td>
<td>il</td>
</tr>
<tr>
<td>esonerare</td>
<td>esonerare</td>
<td>V</td>
<td>V</td>
<td>mod=f</td>
<td>direttiva</td>
<td>direttiva</td>
<td>S</td>
<td>S</td>
<td>num=s</td>
</tr>
</tbody>
</table>

| 96/61/CE. | 96/61/CE. | S | SP | _ |
Ontology Learning: exemplifying Terminology Extraction

- The multi-layered architecture developed by the ItaliaNLP Lab

**NLP tools**
- Tokenization
- Morphological analysis & PoS-tagging
- Lemmatization

**Extraction of candidate terms**
- Linguistic filters
- Statistical filters

**Contrastive ranking**
- Wrts a top list of open-domain terms
- Wrts a top list of terms from a different regulated domain

**Final list of terms ranked on contrastive score**

**List of candidate single and complex terms ranked on statistical filters’ score**
Ontology Learning: exemplifying
Terminology Extraction

- **Single terms**
  - **Linguistic filters:**
    - nouns, e.g. *impianto* (plant), *direttiva* (directive)
  - **Statistical filters:**
    - frequency of occurrence in the input text

<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>impianto</td>
<td>1,570796318</td>
</tr>
<tr>
<td>amministratore</td>
<td>1,570796316</td>
</tr>
<tr>
<td>emissione</td>
<td>1,570796316</td>
</tr>
<tr>
<td>gas</td>
<td>1,570796316</td>
</tr>
<tr>
<td>sostanza</td>
<td>1,570796316</td>
</tr>
<tr>
<td>energia</td>
<td>1,570796316</td>
</tr>
<tr>
<td>serra</td>
<td>1,570796313</td>
</tr>
<tr>
<td>produzione</td>
<td>1,570796312</td>
</tr>
<tr>
<td>deposito</td>
<td>1,570796308</td>
</tr>
<tr>
<td>tabella</td>
<td>1,570796306</td>
</tr>
<tr>
<td>riduzione</td>
<td>1,570796305</td>
</tr>
<tr>
<td>stoccaggio</td>
<td>1,570796304</td>
</tr>
<tr>
<td>veicolo</td>
<td>1,570796304</td>
</tr>
<tr>
<td>quota</td>
<td>1,5707963</td>
</tr>
<tr>
<td>protocollo</td>
<td>1,5707963</td>
</tr>
<tr>
<td>fonte</td>
<td>1,570796297</td>
</tr>
<tr>
<td>costruttore</td>
<td>1,570796297</td>
</tr>
<tr>
<td>elettricità</td>
<td>1,570796297</td>
</tr>
<tr>
<td>inquinamento</td>
<td>1,570796297</td>
</tr>
<tr>
<td>autovettura</td>
<td>1,570796295</td>
</tr>
<tr>
<td>aria</td>
<td>1,570796294</td>
</tr>
<tr>
<td>strategia</td>
<td>1,57079629</td>
</tr>
<tr>
<td>unità</td>
<td>1,570796289</td>
</tr>
<tr>
<td>carbonio</td>
<td>1,570796289</td>
</tr>
<tr>
<td>quantità</td>
<td>1,570796288</td>
</tr>
<tr>
<td>acqua</td>
<td>1,570796287</td>
</tr>
<tr>
<td>gestore</td>
<td>1,570796285</td>
</tr>
<tr>
<td>misurazione</td>
<td>1,570796285</td>
</tr>
<tr>
<td>conte</td>
<td>1,570796284</td>
</tr>
<tr>
<td>trasporto</td>
<td>1,570796283</td>
</tr>
</tbody>
</table>

Corpus of European directives in the environmental domain (Bonin et al., 2010)
Ontology Learning: exemplifying Terminology Extraction

- **Single terms**
  - Linguistic filters:
    - nouns, e.g. *impianto* (plant), *direttiva* (directive)
  - Statistical filters:
    - frequency of occurrence in the input text

- **Multi-word terms**
  - Linguistic filters:
    - noun+preposition+noun, e.g. *riduzione di emissione* (emission reduction); noun+adjective (S+A), e.g. *piano nazionale* (national plan), *normativa comunitaria* (Community legislation)
  - Statistical filters:
    - C-NC Value (Frantzi & Ananiadou 1999), assessing the likelihood for a term of being a well-formed and relevant multi-word term

*Corpus of European directives in the environmental domain (Bonin et al., 2010)*
Ontology Learning: exemplifying Terminology Extraction

Input text → NLP tools → Extraction of candidate terms

- Linguistic filters
- Statistical filters

Output of the statistical filters:

*Open domain terms*, **legal domain terms**, **domain-specific terms** (belonging to the environmental domain) are mixed

Ranking of statistical filters:
- autorità competente 236.120380272
- riferimento al presente direttivo 113.11778156
- destinatario di presente direttivo 108.219717591
- valore limite di emissione 103.436822534
- destinatario di presente decisione 87.2457638653
- limite di emissione 86.9062873351
- sostanza pericoloso 84.8930693328
- giorno successivo 37.5790064648
- anno precedente 23.934467506
- danno ambientale 37.4660023032
Ontology Learning: exemplifying Terminology Extraction

- The multi-layered architecture developed by the ItaliaNLP Lab

![Diagram of NLP tools and terminology extraction process]

- **Input text**
- **NLP tools**
  - Tokenization
  - Morphological analysis & PoS-tagging
  - Lemmatization

- **Extraction of candidate terms**
  - Linguistic filters
  - Statistical filters

- **Contrastive ranking**
  - Wrt a top list of open-domain terms
  - Wrt a top list of terms from a different regulated domain

- **Final list of terms** ranked on contrastive score

- **List of candidate single and complex terms ranked on statistical filters’ score**
Ontology Learning: exemplifying Terminology Extraction

Output of the 1st contrastive phase:

*Open domain terms* are pruned, but *legal domain terms, domain-specific terms* (belonging to the environmental domain) are still mixed.

Ranking of statistical filters

- *autorità competente* 236.120380272
- *riferimento al presente direttivo* 113.117778156
- *destinatario di presente direttivo* 108.219717591
- *valore limite di emissione* 103.436822534
- *destinatario di presente decisione* 113.2457638653
- *limite di emissione* 103.436822534
- *sostanza pericoloso* 84.8903699332
- *giorno successivo* 37.5790064648
- *anno precedente* 23.934467506
- *danno ambientale* 37.4660023032

Contrastive ranking

1st contrastive phase

- *valore limite* 1.57079632502
- *destinatario di presente* 1.57079632361
- *limite di emissione* 1.57079632309
- *valore limite di emissione* 1.57079632286
- *sostanza pericoloso* 1.57079632218
- *aria ambiente* 1.57079632135
- *riferimento al presente direttivo* 1.57079632044
- *autorità competente* 1.57079632041
- *destinatario di presente direttivo* 1.57079631994

Contrast against a top list of terms from a general language corpus (newspaper)
Ontology Learning: exemplifying Terminology Extraction

Output of the 2nd contrastive phase:

**legal domain terms** are singled out by **domain-specific terms** (belonging to the environmental domain)

---

Final term list (2nd contrastive phase)
- sostanza pericoloso 1.57079625565
- salute umano 1.57079624903
- sviluppo sostenibile 1.57079623794
- principio attivo 1.57079622006
- inquinamento atmosferico 1.57079621766
- ................
- norma nazionale 1.57079084047
- testo di disposizione 1.57078547573
- testo di disposizione essenziale 1.57078274091
- disposizione nazionale 1.57078159756
- funzionamento di mercato interno 1.57079632044

---

1st contrastive phase
- valore limite 1.57079632502
- destinatario di presente 1.57079632361
- limite di emissione 1.57079632309
- valore limite di emissione 1.57079632286
- sostanza pericoloso 1.57079632218
- aria ambiente 1.57079632135
- riferimento al presente direttivo 1.57079632044
- autorità competente 1.57079632041
- destinatario di presente direttivo 1.57079631994

Contrastive ranking

Contrast against a top list of terms from a general language corpus (newspaper)

Contrast against a top list of terms from a corpus of European directives regulating a different domain (consumer protection)
Ontology Learning: using extracted terminology to build a legal ontology

- The DALOS (Drafting Legislation with Ontology–based Support) European project (Agnoloni et al., 2009)
  - Aimed at
    - providing law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting
    - enhancing accessibility and alignment of legislation at European level

- Architecture of the DALOS Knowledge Organization System (DALOS ontology)
  - the **Ontological layer**, containing the conceptual modelling at a language independent level
  - the **Lexical layer**, containing multi-lingual terminology conveying the concepts represented at the Ontological layer
Ontology Learning: using extracted terminology to build a legal ontology

- The DALOS (*Drafting Legislation with Ontology–based Support*) project

- **Lexical layer**
  - Terms are
    - automatically extracted from a corpus of Consumer Protection laws
    - automatically organized into taxonomical structures
    - linked to their translation equivalent

- **Ontological layer**
  - Domain-specific concepts and their relationships manually defined by domain experts
Legal Knowledge Extraction: focus on …

- **Identification, extraction and structuring of domain-relevant knowledge**
  - Goal: constructing semantic resources such as domain-specific ontologies or lexicons

- **Semantic annotation** of legal texts
  - Goal: content-based access and querying
Semantic annotation of legal texts: towards a virtuous circle

Textual content (implicit knowledge)

Dynamic content structuring

Linguistic annotation

Knowledge extraction

Structured knowledge (explicit knowledge)

Incremental process of annotation–acquisition–annotation: knowledge acquired from linguistically–annotated texts is projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted

Semantic annotation of legal texts: towards a virtuous circle
Semantic annotation of legal texts: what for

- Tasks requiring NLP-enabled knowledge extraction
  - Legal Argumentation Mining
  - Legal case elements and factors Extraction
  - Legal Text Summarization
  - Court decision Structuring
  - Legal Metadata Extraction
  - Legal definition Extraction
  - Legal citation Extraction
  - Legal Information Retrieval
  - ...
Semantic annotation of legal texts: example (1)

- Legal case elements and factors Extraction for Legal Argumentation Mining
  - Adam Wyner (tomorrow morning)

- NLP tools used to make explicit relevant legal facts and legal roles starting from their linguistic realization in a collection of legal cases
  - E.g. the Appellee, Defendant, Plaintiff, etc.
  - E.g. the Disclosure-in-Negotiation fact (i.e. the fact that the plaintiff disclosed information during negotiation with defendant)

The annotation are the building blocks of a language of formal rules
Semantic annotation of legal texts: example (2)

- **Legal definition Extraction**
  - Walter and Pinkal, 2006: from German court decisions

- NLP tools are used to identify legal definitions on the basis of the linguistic realization of *definiendum* and *definiens*
  - “One-family row-houses have insufficient noise insulation if the separating wall is one-layered”

- The linguistic structure is transformed to a semantic representation by a series of heuristic rules

- Promising step for Ontology Learning purposes
Legal Metadata Extraction

- Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - It combines
    - a set of tools which transform a plain text in XML, detect references and classify provisions (i.e. `xmLeges` tools)
    - a suite of NLP tools for the analysis of Italian texts
  - It aims at supporting the consolidation of legislative texts process (in force law)
  - It provides a formalized representation of textual amendments by a metadata set
    - Repeal, substitution and integration
  - The text modification is performed on the metadata interpretation
## Semantic annotation of legal texts: example (3)

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pos</td>
<td>information on the amending provision</td>
</tr>
<tr>
<td>pos:xlink</td>
<td>ID reference to the amending provision</td>
</tr>
<tr>
<td>norm</td>
<td>information on the norm to be amended</td>
</tr>
<tr>
<td>norm:xlink</td>
<td>URN reference to the norm</td>
</tr>
<tr>
<td>- pos</td>
<td>further information on the norm</td>
</tr>
<tr>
<td>- pos:xlink</td>
<td>URN reference to the norm with the partition ID</td>
</tr>
<tr>
<td>-- border</td>
<td>information on further narrower container</td>
</tr>
<tr>
<td>-- border:type</td>
<td>container type (e.g. point, “alinea”, period, etc.)</td>
</tr>
<tr>
<td>-- border:num</td>
<td>container label expressed by a number or a letter</td>
</tr>
<tr>
<td>-- border:ord</td>
<td>container position expressed by an ordinal (e.g. 2nd) or a relative (e.g. last) number</td>
</tr>
<tr>
<td>position</td>
<td>information on the specific modifying point within the narrowest container</td>
</tr>
<tr>
<td>- pos</td>
<td>information on a string (quoted) and a bound of the deleting or inserting</td>
</tr>
<tr>
<td>- pos:xlink</td>
<td>ID reference to the string, a bound of which is the beginning of the modifying text</td>
</tr>
<tr>
<td>- pos:where</td>
<td>specific bound of the string or container (before, after, start, end)</td>
</tr>
</tbody>
</table>

| novellando    | information on the outgoing text                                             |
| - type        | “novellando” type (e.g. article, paragraph, “alinea”, period, words, etc.)   |
| - type:value  | information on the outgoing string (in quotes)                              |
| - pos         | ID reference to the string that is either the outgoing text, or the beginning or ending of the outgoing text |
| - pos:xlink   | information on the meaning of the string                                    |
| - role        | string role: beginning (from) or ending (up to) of the outgoing text         |
| - role:value  | information on the incoming text                                            |
| novella       | information on the incoming text                                            |
| - type        | “novella” type (e.g. article, paragraph, “alinea”, period, words, etc.)      |
| - type:value  | information on the incoming string (quoted)                                 |
| - pos         | ID reference to the incoming string                                         |
Legal Metadata Extraction

Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG

An example

“All’articolo 1, comma 1, della legge 8 febbraio 2001, n. 12, la lettera d) è abrogata” (In article 1, paragraph 1, of the act 8 February 2001, n. 12, letter d) is repealed)
Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (Metadata Extraction from Legal Texts) system jointly developed by ILC and ITTIG
  - An example
    - “All’REF mod31-rif2#art1-com1, la lettera d) è abrogata” (In REF mod31-rif2#art1-com1, letter d) is repealed)
Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - An example
    - The sentence was linguistically analyzed at a shallow syntactic level of analysis

```
[[CC:P_C][PREP:A#E][DET:LO#RD@MS][POTGOV:REF#SP@NN]]
[[CC:U_C][FORM:MOD31-RIF1#ART1-COM1]]
[[CC:PUNC_C][PUNCTYPE: ,@]]
[[CC:N_C][DET:LO#RD@FS][AGR:@FS][POTGOV:LETTERA#S@FS]]
[[CC:N_C][AGR:@FS][POTGOV:D#S@FS]]
[[CC:PUNC_C][PUNCTYPE: )#@]]
[[CC:FV_C][AUX:ESSERE#V@S3IP][POTGOV:ABROGARE#V@FSPR]]
[[CC:PUNC_C][PUNCTYPE: .#@]]
```
Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - An example
    - The annotation of informative metadata was carried out by a finite-state compiler which uses a specialized grammar covering the amendment types considered on the basis of patterns formalized in terms of regular expressions operating over sequences of chunks

```xml
<repeal>
  In <norm>REF mod31-rif1#art1-com1#</norm>, <border>
    <border:type>letter</border:type>
    <border:num>d</border:num>
  </border> is repealed.
</repeal>
```
Natural Language Processing techniques represent a key ingredient for Legal Knowledge Management.

Knowledge Creation: Legal Ontologies and Lexicons

Natural Language Processing tools

Texts

Structured knowledge

Knowledge Use: “Intelligent” content access

Hopefully, thanks to NLP Legal Search Engines will be able to access the content embedded in texts more effectively.
One of the main obstacles to progress in the field of artificial intelligence and law is the **natural language barrier**.

L. Thorne McCarty, International Conference on AI and Law (ICAIL-2007)

**Natural Language Processing** combined with **Knowledge Extraction** techniques can help removing or at least penetrating the natural language barrier in the AI&Law field.
The NLP tools and techniques have been developed in the framework of the activities of the people of ItaliaNLP Lab at the Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)

http://www.italianlp.it/

Special thanks to Felice Dell’Orletta
On-line demos

- Linguistic analysis of Italian and English texts

- Term extraction from Italian and English texts
References

Ontology Learning


Ontology Learning in the legal domain (1)

References

- **Ontology Learning in the legal domain (2)**
  - Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds.), 2010, SEMANTIC PROCESSING OF LEGAL TEXTS, LNCS 6036, Springer

- PART II - Legal Text Processing and Construction of Knowledge Resources
  - Automatic Identification of Legal Terms in Czech Law Texts by Karel Pala, Pavel Rychlý, Pavel Šmerk
  - Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain by Enrico Francesconi, Simonetta Montemagni, Wim Peters, Daniela Tiscornia
  - Ontology Based Law Discovery by Alessio Bosca, Luca Dini
  - Multilevel Legal Ontologies by Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P. Radicioni, Piercarlo Rossi

References

- **Legal Language Processing (computational linguistics literature)**
  - Alessandro Mazzei, Cristina Bosco (2012), Simple Parser Combination. In Proceedings of the 4th Workshop on “Semantic Processing of Legal Texts”
  - Venturi G. (2008), Parsing Legal texts. A contrastive study with a View to Knowledge Management Applications. In Proceedings of the “Ist Workshop Semantic Processing of Legal Texts”, held in conjunction with LREC, Marrakech, Morocco, May 26–1,
References

Semantic Processing of legal texts

- Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds.), 2010, SEMANTIC PROCESSING OF LEGAL TEXTS, LNCS 6036, Springer