From text to knowledge: the starting point

- 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
From text to knowledge: the starting point

- 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

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)
- 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**
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

The need for increasingly sophisticated applications based on Natural Language Processing technologies and aimed at overcoming the knowledge acquisition bottleneck
Summary

- Natural Language Processing tools
  - What and what for
  - An example

- From text to knowledge
  - The general approach
  - The main challenges of the legal domain

- Legal Knowledge Extraction
  - What and what for
  - An example
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 allow
  - accessing the **linguistic** and **domain-specific** knowledge contained in texts
  - structuring the textual content
The Natural Language Processing tools: what

- Tools that enable computers to derive meaning from human or natural language input

- They are a component of artificial intelligence, computer science and linguistics concerned with processing texts and making information accessible to computer applications

- They make machine-readable the linguistic structure implicitly embedded in texts
  - Automatic linguistic annotation process
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.).
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>
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
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>
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.)
The linguistic annotation tools @ ItaliaNLP Lab

- **LinguA** is a state-of-the-art linguistic annotation pipeline which combines rule-based and machine learning algorithms
  - developed by ILC and the University of Pisa

- Morpho-syntactic annotation (PoS tagger developed by Dell’Orletta, 2009)
  - Evalita 2009: accuracy = 96.34%
  - State-of-the-art for Italian

- Dependency syntactic annotation (DeSR parser, Attardi & Dell’Orletta, 2009)
  - Conll-2007: 81.3% LAS
  - Evalita 2009: 83.38% LAS
  - State-of-the-art for Italian

Demo at http://www.italianlp.it/demo/linguistic-annotation-tool/
LinguA is a linguistic annotation pipeline which combines rule-based and ML algorithms.
Linguistic Annotation Pipeline

Thanks to Stefano Del Rossi for the dependency graph visualization.
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
From text to knowledge: the general approach

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.

Textual content (implicit knowledge)

Structured knowledge (explicit knowledge)

Dynamic content structuring

Knowledge extraction

Linguistic annotation
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 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 and its impact on NLP tools

- Legal texts differ significantly with respect to ordinary language texts (e.g. newspapers)
  - typically correlated with syntactic complexity

- Dramatic drop of accuracy when NLP tools are tested on domains outside of the data from which they are trained or developed on (Gildea, 2001)
  - Key role of natural language syntactic parsing which represents a prerequisite for any advanced legal text processing task

- What is the performance of state-of-the-art NLP tools on legal texts?
  - A key issue for all NLP-based Legal Knowledge Extraction tasks
The peculiarity of legal language and its impact on NLP tools

- Recently, two initiatives 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 initiatives:
  - Domain Adaptation Track at Evalita 2011 – Italian
  - SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts – Italian and English
## The peculiarity of legal language and its impact on NLP tools

The Evalita 2011 results for dependency parsing

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>Performance</th>
<th>Performance after Domain Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspaper</td>
<td>Newspaper</td>
<td>82.09% Labelled Attachment Score (LAS)</td>
<td>---</td>
</tr>
<tr>
<td>Newspaper</td>
<td>Legal texts</td>
<td>75.85% LAS</td>
<td>80.83% LAS</td>
</tr>
</tbody>
</table>

- 6.24 %

+ 5 %
From text to knowledge: the main challenges of the legal domain

- **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
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:
- 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 can be a helpful starting point for any further construction of domain-specific knowledge base where domain-relevant and the specific domain knowledge is kept separate
- It is closely related to the reusability and interoperability issue
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)
**T2K** (Text-To-Knowledge) combines a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning which are dynamically integrated to provide an accurate representation of the domain-specific context of text corpora in different domains (Dell’Orletta et al., 2014)

**T2K system**

- **Linguistic pre-processing**
  - Linguistic Analysis Tools:
    - Sentence Splitter
    - Tokenizer
    - Morphological Analyzer
    - PoS Tagger
    - Chunker
    - Dependency Parser

- **Knowledge extraction**
  - Information Extraction Tools:
    - Domain-specific Entities extractor
    - Named Entity tagger
    - Relation extractor
  - Knowledge Graph Tools:
    - Semantic annotator
    - Indexer
    - Graph creator
    - Graph Visualizer

**Semantic Annotation**

**Knowledge graph**

**Index of Content**
In T2K the NLP tools were trained on two training sets: the ISST-TANL treebank consisting of newspaper articles and a syntactically annotated corpus of Italian legislative and administrative texts.
From text to knowledge @ ItaliaNLP Lab

- The IE tools allow extracting
  - domain-specific entities (Bonin et al. 2010)
    - e.g. nominal terminology, verbs (both single- and multi-word expressions)
  - Named entities
    - i.e. Person, Location, Organization and Geopolitical
  - relations between the extracted entities
    - taxonomical
      - e.g. health research, international research, cancer research or research projects, research infrastructure
    - co-occurrence within the same context and similarity on the basis of shared contexts

- They result in
  - multi-dimensional knowledge representation graph
  - document collection index and semantic annotation
Terminology Extraction
**Input corpus:** a collection of European Italian Directives on consumer protection

**TERMINOLOGY EXTRACTION**

<table>
<thead>
<tr>
<th>Lemma of Term</th>
<th>Domain Relevance</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>trade service</td>
<td>100.00%</td>
<td>421</td>
</tr>
<tr>
<td>Sub-contracted operation</td>
<td>100.00%</td>
<td>398</td>
</tr>
<tr>
<td>custom authority</td>
<td>100.00%</td>
<td>326</td>
</tr>
<tr>
<td>wholesale trade service</td>
<td>100.00%</td>
<td>154</td>
</tr>
<tr>
<td>personal data</td>
<td>100.00%</td>
<td>151</td>
</tr>
<tr>
<td>contract basis</td>
<td>100.00%</td>
<td>120</td>
</tr>
<tr>
<td>third country</td>
<td>100.00%</td>
<td>117</td>
</tr>
<tr>
<td>convention</td>
<td>100.00%</td>
<td>584</td>
</tr>
<tr>
<td>national law</td>
<td>100.00%</td>
<td>112</td>
</tr>
<tr>
<td>tariff information</td>
<td>100.00%</td>
<td>102</td>
</tr>
<tr>
<td>custom debt</td>
<td>100.00%</td>
<td>100</td>
</tr>
<tr>
<td>travel document</td>
<td>100.00%</td>
<td>86</td>
</tr>
<tr>
<td>import duty</td>
<td>100.00%</td>
<td>82</td>
</tr>
<tr>
<td>free circulation</td>
<td>100.00%</td>
<td>82</td>
</tr>
</tbody>
</table>
T2K handles this challenge thanks to

a **multilayered contrastive approach** to entity extraction

- The domain relevance of entities is assessed on the basis of the contrastive distribution of relevant candidate entities across an input corpus and a different corpus.
- The contrastive analysis is iterated twice:
  - against a top list of open-domain entities (e.g. from newspapers) to prune common entities (e.g. *following day*).
  - against a top list of entities from e.g. a different regulated domain to discriminate legal and regulated-domain entities.

- **a new term ranking function** suitable for handling variation in low frequency events
  - E.g. in the legal texts, regulated-domain entities have low frequency and they are sparse.
Named Entity Extraction
Input corpus: a collection of European Italian Directives on consumer protection
Relation Extraction
**Input corpus:** a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

**Relation Extraction**

E.g.: terms in relation with *imaging cerebrale* in criminal case law texts

<table>
<thead>
<tr>
<th>imaging cerebrale <em>(brain imaging)</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>genetica molecolare <em>(molecular genetics)</em></td>
</tr>
<tr>
<td>difesa <em>(defense)</em></td>
</tr>
<tr>
<td>valutazione <em>(evaluation)</em></td>
</tr>
<tr>
<td>colloqui clinici <em>(clinical interviews)</em></td>
</tr>
<tr>
<td>emergenze psichiatriche <em>(psychiatric emergencies)</em></td>
</tr>
<tr>
<td>accertamenti psichiatrici <em>(psychiatric inspections)</em></td>
</tr>
</tbody>
</table>
**Input corpus:** a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

---

**Relation Extraction**

E.g.: terms in relation with *imaging cerebrale* in criminal case law texts

<table>
<thead>
<tr>
<th>Imaging cerebrale (<em>brain imaging</em>)</th>
<th>Genetica molecolare (<em>molecular genetics</em>)</th>
<th>Quadro clinico (<em>medical case</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difesa (<em>defense</em>)</td>
<td>Comportamenti illeciti (<em>illegal behaviours</em>)</td>
<td></td>
</tr>
<tr>
<td>Valutazione (<em>evaluation</em>)</td>
<td>Nesso causale (<em>causal relationship</em>)</td>
<td></td>
</tr>
<tr>
<td>Colloqui clinici (<em>clinical interviews</em>)</td>
<td>Apporto tecnico (<em>technical contribution</em>)</td>
<td></td>
</tr>
<tr>
<td>Emergenze psichiatriche (<em>psychiatric emergencies</em>)</td>
<td>Sfera psichica (<em>psychic sphere</em>)</td>
<td></td>
</tr>
<tr>
<td>Accertamenti psichiatrici (<em>psychiatric inspections</em>)</td>
<td>Imputata (<em>defendant</em>)</td>
<td></td>
</tr>
</tbody>
</table>

---

Sia le emergenze psichiatriche, completate dalle risultanze dell' *imaging cerebrale* e di *genetica molecolare*, che quelle processuali consentono di rilevare gravi segni di disfunzionalità psichica, eterogenei ma convergenti nell' indicare un nesso causale tra i disturbi dell' *imputata* ed i suoi comportamenti illeciti.
From text to knowledge: the general approach

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.

- **Textual content** (implicit knowledge)

- **Structured knowledge** (explicit knowledge)

- **Dynamic content structuring**

- **Linguistic annotation**

- **Knowledge extraction**
Document indexing

System

Knowledge extraction

Information Extraction
- Tools
  - Domain-specific Entities extractor
  - Named Entity tagger
  - Relation extractor

Knowledge Graph
- Tools
  - Semantic annotator
  - Indexer
  - Graph creator
  - Graph Visualizer

Semantic annotation

Knowledge graph

Index of Content
The acquired knowledge (e.g. terms, named entities) is used for document indexing on the basis of the extracted domain-specific knowledge.

Input corpus: a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms.

<table>
<thead>
<tr>
<th>Term</th>
<th>Document</th>
<th>TF*IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>risonanza magnetica</td>
<td>Penale/merito/merito/massime/Cort_Assis_App.Trieste, pen., mass., 01-10-2009.txt</td>
<td>0.141522742511</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_I, Sent., (ud_25-10-2012) 21-11-2012, n_45559.txt</td>
<td>0.0637848980331</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_I, Sent., (ud_16-12-2011) 03-05-2012, n_16281.txt</td>
<td>0.0576907994949</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2010/Cass_pen_Sez_IV, (ud_20-11-2009) 14-01-2010, n_1489.txt</td>
<td>0.0485220831466</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2006/Cass_pen_Sez_III, (ud_21-06-2006) 10-10-2006, n_33974.txt</td>
<td>0.0358002194494</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2007/Cass_pen_Sez_I, Sent., (ud_13-12-2006) 02-03-2007, n_9173.txt</td>
<td>0.0334223450948</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/merito/merito/2006/Trib_Ord., Genova, pen., Sez_III, 21-07-2006.txt</td>
<td>0.0301915184023</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2006/Cass_pen_Sez_V, Sent., (ud_18-05-2006) 14-11-2006, n_37452.txt</td>
<td>0.0190026279006</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2013/Cass_pen_Sez_III, Sent., (ud_04-12-2012) 22-01-2013, n_3258.txt</td>
<td>0.0174854353681</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_I, Sent., (ud_03-07-2012) 20-07-2012, n_29707.txt</td>
<td>0.0159462245083</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/merito/merito/2013/Trib.Ord., Venezia, pen., Sez., 08-04-2013_utf8.txt</td>
<td>0.015652284866</td>
</tr>
<tr>
<td>risonanza magnetica</td>
<td>Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_IV, (ud_01-12-2011) 31-01-2012, n_3986.txt</td>
<td>0.0147036615596</td>
</tr>
</tbody>
</table>

The term *risonanza magnetica* (magnetic resonance) occurs both in lower courts and in the Court of Cassation but it is ‘more relevant’ in the criminal case resolved by the Trieste ordinary tribunal in 2009.
Semantic annotation
La sentenza ritiene azionato, pur in assenza di espressa qualificazione in tal senso nell’atto introduttivo del giudizio, il diritto al risarcimento del danno, ex art. 2043 c.c., per violazione dell’obbligo dello Stato di dare attuazione alle direttive comunitarie che imponevano di remunerare adeguatamente il medico per la frequenza di un corso di specializzazione; considera comprovato, in assenza di contestazioni specifiche, che il C. avesse superato il corso di formazione quadriennale, come da attestazione del 5.11.1992, con frequenza a tempo pieno e senza svolgimento di attività libero-professionale; dichiara inammissibile l’eccezione di prescrizione quinquennale sollevata dall’amministrazione ed accolta dal primo giudice, sul rilievo che era stata formulata, senza le necessarie allegazioni in fatto e diritto, con riferimento all’art. 2948 c.c., n. 4, in termini, quindi, non pertinenti al rapporto giuridico dedotto in giudizio, atteso che non si trattava di rapporto di impiego pubblico (prospettazione su cui si fondava il difetto di giurisdizione ordinaria, eccepito dall’amministrazione in primo grado) e di responsabilità contrattuale; liquida il risarcimento nell’importo di L. 13.000.000 annue (Euro 6.713,93) secondo il parametro fornito dalla L. n. 370 del 1999, art. 1, comma 1 (borsa di studio annuale per i medici ammessi presso le università alle scuole di specializzazione in medicina dall’anno accademico 1983-1984 all’anno accademico 1990-1991, in attuazione di giudicati amministrativi), con l’aggiunta della rivalutazione monetaria e degli interessi legali dalla maturazione del credito, fissata alla data del 5 novembre 1992.
The semantically annotated corpus can be used by a search engine to retrieve the text spans containing the information searched for.

Example: prescrizione quinquennale
Knowledge graph
**Knowledge graph**

**Input corpus:** a collection of European Italian Directives on consumer protection

In T2K the extracted information interact resulting in a multidimensional knowledge representation graph creating the prerequisites for sophisticated text mining processes.
Knowledge graph

**Input corpus:** a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

The sub-graph of *imaging cerebrale* in criminal case law texts
To sum up: from bricks of knowledge to a domain ontology
To sum up: from bricks of knowledge to a domain ontology

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)
Approaches to Ontology Design and Development: top-down vs bottom-up

**TOP-DOWN**

ontology construction starts by modeling top level concepts, which are then subsequently refined.

this approach is typically carried out manually by domain experts and leads to a high-quality engineered ontology.
Approaches to Ontology Design and Development: top-down vs bottom-up

**TOP-DOWN**

- ontology construction starts by modeling top level concepts, which are then subsequently refined
- this approach is typically carried out manually by domain experts and leads to a high-quality engineered ontology

**BOTTOM-UP**

- it starts from the assumption that most concepts and conceptual structures of the domain are contained in documents
- the terminological and conceptual knowledge contained in document collections is semi-automatically extracted from texts, thus creating the basis for ontology construction
Ontology Learning: an incremental process

- The various steps of Ontology Learning from texts can be arranged in a “layer cake” of increasingly complex subtasks
  - (Buitelaar, Cimiano and Magnini, 2005)

\[ \forall x, y \ (\text{sufferFrom}(x, y) \rightarrow \text{ill}(x)) \]

\text{cure (dom:DOCTOR, range:DISEASE)}

\text{is_a (DOCTOR, PERSON)}

\text{DISEASE:=<Int,Ext,Lex>}

\{ \text{disease, illness} \}

\text{disease, illness, hospital}
Ontology Learning: an example

- 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: an example

- 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
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.
Conclusion

- Natural Language Processing techniques represent a key ingredient for Legal Knowledge Extraction and Management
Natural Language Processing techniques represent a key ingredient for Legal Knowledge Extraction and Management.

Knowledge Creation:
Legal Ontologies and Lexicons

Knowledge Use:
“Intelligent” content access

Hopefully, thanks to NLP Legal Search Engines will be able to access the content embedded in texts more effectively.
Credits

- The NLP tools and techniques have been developed in the framework of the activities of the 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
  - http://www.italianlp.it/demo/linguistic-annotation-tool/

- Text-To-Knowledge (T2K)
  - http://www.italianlp.it/demo/t2k-text-to-knowledge/
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 Francescon, Simonetta Montemagni, Wim Peters, Daniela Tiscornia
    - Ontology Based Law Discovery by Alessio Bosca, Luca Dini
    - Multilevel Legal Ontologies by Gianmario Ajani, Guido Boella, Leonardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P. Radicioni, Piercarlo Rossi
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 “1st 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