Ontology-driven Generation of Training Paths in the Legal Domain

http://dx.doi.org/10.3991/ijet.v10i7.4609

Nicola Capuano, Andrea Longhi, Saverio Salerno, Daniele Toti
University of Salerno, Italy

***Abstract*—This paper presents a methodology for helping citizens obtain guidance and training when submitting a natural language description of a legal case they are interested in. This is done via an automatic mechanism, which firstly extracts relevant legal concepts from the given textual description, by relying upon an underlying legal ontology built for such a purpose and an enrichment process based on common-sense knowledge. Then, it proceeds to generate a training path meant to provide citizens with a better understanding of the legal issues arising from the given case, with corresponding links to relevant laws and jurisprudence retrieved from an external legal repository. This work describes the creation of the underlying legal ontology from existing sources and the ontology integration algorithm used for its production; besides, it details the generation of the training paths and reports the results of the preliminary experimentation that has been carried out so far. This methodology has been implemented in an Online Dispute Resolution (ODR) system that is part of an Italian initiative for assisted legal mediation.**

***Index Terms*—adaptive learning systems; semantic search; online dispute resolution; text analysis; knowledge representation; ontology engineering; ontology integration**

I. INTRODUCTION

Within the legal context, mediation is a form of alternative dispute resolution whose purpose is to try and resolve disputes between two or more parties, by resorting to the help of a "mediator" meant to support parties in order for them to reach a mutual agreement. In Italy, the first modern mediation statutes were issued in 1993, even though the practice of mediation did not take real hold until 2011, when the Italian Government introduced a mandatory pretrial mediation for civil and commercial cases; a revised regulation was later enforced in 2013. Since then, a certain degree of improvement has been detected when mediation is applied, in terms of reduced time to reach a settlement in litigations and consequently an increase in the overall efficiency of the legal procedures. Nevertheless, awareness for the benefits of mediation still needs to be spread among citizens, and effective tools to fully support it are sorely needed.

In this regard, the eJRM project[1], which stands for "electronic Justice Relationship Management" and is supported by the Italian Ministry of University and Research, was born for defining, implementing and experimenting innovative methodologies and technologies for online mediations. The major expected result of this project is a full-fledged system for helping citizens and mediators begin and carry out a mediation, via a number of online, collaborative tools for enabling a remote communication and resolution among parties, in full accordance with all of the provisions of the mediation statutes.

Aside from that, the project is meant to provide its users with the possibility of formulating a case in natural language or via a structured interview [1], and let the system assist them by generating a custom-tailored Knowledge Path (KP), where the involved legal concepts are outlined and the corresponding information is displayed, by integrating training modules based on *story-telling* [2]. This enables citizens to autonomously perform a preliminary assessment of their legal case and decide the course of action most suitable to their needs to be subsequently taken (e.g. suing the other party, initiating mediation or giving up).

In this paper, the process related to the generation of the aforementioned KPs within the eJRM system is described, along with its underlying ontological models and the semantic techniques used; furthermore, preliminary results coming from its experimentation are reported, as well as some details related to the current prototype implementing this process. A shorter version of this discussion has been published in [3].

The paper is structured as follows. In Section II, related work is discussed. Section III describes the ontological structure that lies at the core of the KP generation process, whereas the latter is detailed in Section IV. Then, in Section V, experimental results are reported and commented. Finally, in Section VI conclusions are drawn.

II. RELATED WORK

Recent developments in knowledge representation and automatic reasoning make it possible, for software systems, to support juridical processes via the development of models of legal knowledge. Semantic technologies applied to this domain are progressively enabling the development of applications to increase the transparency of legal rules, to analyze the impact of changes in laws, to research inconsistencies and non-regulated scenarios, and to create portals for legal advice.

Earlier attempts to use knowledge representation in legal systems date back to a time when semantic technologies were still in their embryo form. Among those, it is possible to cite the *Frame-based Ontology of Law* (FBO) [4] that considers a legal system made of norms, acts and concepts, in contrast with the *FOLaw* ontology [5], adopting a functional perspective that relies on a structure made up of several different kinds of knowledge. In [6] authors also proposed an approach where law was defined as a dynamic and interconnected system of states of affairs

[1] http://www.ejrm.it

evolving over time. A more recent initiative is the project *Estrella* [8] aimed at defining a *Legal Knowledge Interchange Format* (LKIF), based on RDF and OWL, for the representation of legal concepts. LKIF is made of about 200 concepts that can be used as a starting point to define a legal system. LKIF lies at the core of *HARNESS* [9], a system able to check whether a case complies with specific legal rules or not.

There are several other projects involving the use of semantic technologies in the legal field. One is *ICT4Law*, where a *Legal Taxonomy Syllabus* [10] has been defined to represent legal information at different levels, distinguishing between terms and their Interlingua meanings; according to European Directives, the syllabus is able to align specific legal terms of each country with a core ontology of legal concepts. Another project is *Ontomedia* [11] that deals with the use of ontologies for online mediation. The project defined a specific *Mediation Core Ontology* (MCO), based on OWL, for the semantic representation of legal documents acquired during the mediation steps. The developed system, based on MCO, provides advanced retrieval features as well as knowledge-based tools for the formal definition of new mediation cases.

LOIS [12], instead, is a multilingual lexical resource in the legal field based on *WordNet* including about 35.000 concepts in five European languages. The related *DALOS* project [13] has built an ontology-linguistic resource, based on LOIS, to be used in EU legislative drafting process as well as in the national transpositions of EU Directives. The *MetaSearch* project [7], on the other hand, aims at developing a system for the search, indexing and automatic mark-up of legal documents enabling semantic search and retrieval to be performed by users. In order to do that, legal ontologies, specific relevance detection algorithms and a system for automatic mark-up and indexing are employed.

Such systems, like many others that try to introduce semantic technologies in juridical processes [14], are mainly targeted to support legal specialists in juridical tasks like the formal definition and the legal assessment of a case, the semantic retrieval and the alignment of legal documents, and so forth. Few attempts have been made so far, to the best of the authors' knowledge, to use such technologies to support citizens (with few or no juridical background) in obtaining guidance and training on legal topics. In fact, it is the authors' belief that common citizens should not be forced to use complex semantic tools as well as formal languages to describe the case on which they would be supported; instead, it would be better for them to take advantage, for this purpose, of user-friendly approaches like natural language or structured interviews.

In the event of adopting natural language, the "language gap" problem must be also taken into account. Common citizens are not used to employing juridical language and, rather often, they rely on non-appropriate terms to describe legal concepts.

Besides, providing guidance to the common citizen does not simply mean finding relevant resources with respect to the expressed case (like in traditional or semantics-based information retrieval): this also requires a feasible training path to interconnect the retrieved resources, so that it may be possible to transfer, in the most effective way, the relevant knowledge behind them to the given legal case.

The system described in this paper follows this specific direction, since it strives to directly provide citizens with innovative features, by integrating semantics and ontology-based approaches with models and techniques coming from adaptive learning systems [15]; furthermore, a methodology based on common-sense knowledge is also used to overcome the "language gap" issue earlier mentioned.

III. THE LEGAL ONTOLOGY

As the underlying foundation of the algorithms and techniques that have been devised, which will be described later in the paper, lies a model able to formally describe legal concepts and related training and informative resources. Such a model is composed of three abstraction levels, as depicted in Figure 1.

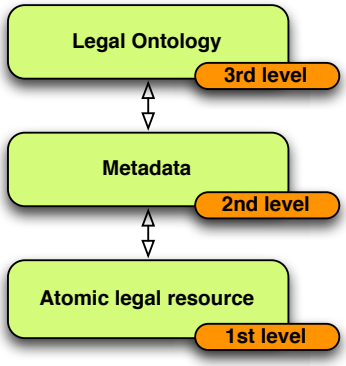

Figure 1. Knowledge management abstraction levels.

The lower level consists of *atomic legal resources* that are used to build KPs (e.g. a legal principle, a precept, a sentence, a learning object etc.). The intermediate level is made of *metadata* describing the legal resources via a set of attributes that may vary according to the scope (educational or informational) of the resource. The higher level, named *Legal Ontology*, deals with the conceptual management of the available resources.

The *Legal Ontology* is a structure composed of concepts and relationships between concepts. More formally it can be defined as a graph O (C, R_1,… R_n) where C is a set of nodes that represent the concepts and each R_i is a set of edges which correspond to a specific type of relationship. Two sets of relationship types are defined, the first (informative) is purposed to define a structured dictionary of legal terms, the second (educational) is aimed at introducing useful properties for training.

The *informative set* includes the following relationships, compliant to the SKOS[2] specification.

a) *NT* (a, b) means that the concept a is a *narrower term* with respect to the concept b, i.e. a has a more specific meaning that b;

b) *BT* (a, b) means that the concept a is a *broader term* with respect to the concept b, i.e. a has a more general meaning that b;

c) *RT* (a, b) means that the concept a is generally *related to* the concept b.

The *educational* set includes the following relationships in accordance with [16]:

[2] http://www.w3.org/2004/02/skos/

a) *HP* (*a*, *b*) means that the concept *b is part of a*, i.e. *a* is understood if and only if every *b* so that *b* is part of *a* is understood;
b) *IRB* (*a*, *b*) means that the concept a *is required by* the concept *b*, i.e. a necessary condition to study *b* is to have understood *a* before;
c) *SO* (*a*, *b*) means that the *suggested order* between the two concepts is that *a* precedes *b*, i.e. to favor learning, it is desirable to study *a* before *b*.

In addition, to each node $c \in C$ several information is connected: a name label $N(c)$, a textual description $D(c)$ and a weighted set of terms $T(c)$ characterizing the concept. This latter set, generated by the *enrichment* process described in III.C, has the purpose of adding commonsense meaning to each concept to limit the "language gap" issue earlier described. The subsequent subsections will respectively describe how concepts belonging to the Legal Ontology were defined (A), how the Ontology was built by integrating such concepts (B), how additional metadata have been added (C) and how the ontological concepts have been further enriched with common-sense knowledge (D).

A. Definition of ontological concepts

In order to fill the defined ontological structure with concepts and relationships, the first step was to investigate existing legal ontologies like those mentioned in Section II. Such ontologies, however, turned out to be either too application-specific, and thus severely restricting their general applicability, or too abstract, dealing with general concepts like "norm", "rule", "legal action" etc. without their specific instances. The legal ontology to be defined had instead the purpose of supporting the classification of legal topics rather than the definition of legal abstractions: as such, for building it, it was deemed necessary to collect and integrate information coming from two major legal sources.

One of them is *EuroVoc*[3]: a multilingual thesaurus defined by the European Community (EC) and built according to the SKOS formalism for the classification of directives, laws and treaties. The current version of *EuroVoc* (as of November 2014) includes 6.892 concepts and is available in 23 official EU languages. To each concept a textual description is attached as well as a list of aliases; the concepts are mutually interlinked via hierarchical relationships like *BT* (broader term) and *NT* (narrower term), as well as associative relationships like *RT* (related term). This thesaurus covers all the activity fields of European institutions i.e. politics, international relations, law, economics, trade, finance, social questions, education, science, business, employment, food, environment, agriculture, forestry, production, technology, research, energy, industry, geography.

Choosing *EuroVoc* as one of the legal sources for the system on one hand would have ensured interoperability with many legal databases and repositories, since it is currently adopted by all EC institutions and by European national parliaments; on the other hand, the scope and depth of its included concepts would have fallen short for supporting an effective detection of legal cases from natural language texts, especially for the Italian law. That is why, alongside *EuroVoc*, *ItalGiure*[4] was taken into account: this is one of the major repositories for legal taxonomies in Italy curated by the Italian Court of Cassation, which includes Italian laws and the main corresponding jurisprudence from any law sector classified accordingly. In this regard, given the fact that mediation is usually adopted for civil law cases, only the civil section of the *ItalGiure* repository was considered, consisting of 12701 terms.

EuroVoc has been converted to an ontological representation by importing it as it is, with its *BT*, *NT* and *RT* relationships, whereas *ItalGiure* has been turned to a hierarchical representation by using *NT* and *BT* relationships only, since it did not support associative relationships. Afterwards, an integration of the two representations has been carried out, by means of an ontology integration mechanism specifically devised for this task as described in the following subsection.

B. Ontology integration

The applied ontology integration mechanism is a three-phase process where (0) the input ontologies are indexed and prepared, (1) a matching is performed among concepts from the input ontologies in order to find correspondences among them, and (2) the identified correspondences are turned into *RT* relationships according to the degree and score of the match. The full description of this process is described below.

The initial phase, or Phase 0, takes care of extracting concepts from the input ontologies in terms of their labels, and of indexing the second ontology for enabling the fuzzy full-text search described below while minimizing computational time.

The second phase, or Phase 1, consists of the execution of the matching algorithm defined as follows. Given the *ItalGiure* ontology O_I, the *EuroVoc* ontology O_E, t_p threshold for partial matches and a t_a threshold for approximate matches, $\forall$ concept $c_j \in O_I$:

- a corresponding set of concepts K_E potentially similar to c_j is retrieved from O_E via a fuzzy full-text search; this operation helps us avoid the need of scanning each time the entire list of concepts from O_E, by performing an initial pruning that leaves only the most promising candidates. This search is a combination of several syntactical distances and is entrusted to a specific full-text search tool (see Section V for implementation details).
- $\forall\ k_j \in O_E$:
  - an "exact" match is identified whenever one of the following conditions is met:
    - if $lowercase(c_j) = \text{lowercase}(k_j)$;
    - if $lemma(c_j) = lemma(k_j)$;
    - if c_j and k_j are made up of multiple words (tokens), and one of the above equality conditions applies regardless of permutations of their respective tokens.
  - a "partial" match is identified whenever the conditions for an exact match do not apply and the following condition is met:
    - if c_j and k_j are made up of multiple words (tokens) and one of the equality conditions for the exact matches applies on a number of individual tokens; here, only "relevant" tokens are considered, by removing stop-words, prepositions, connectors and the like. If (#relevant tokens) /

[3] http://eurovoc.europa.eu/
[4] http://www.italgiure.giustizia.it/

(#max number of relevant tokens) = $s \geq t_p$, a partial match is obtained, with s as its score;

- an "approximate" match is identified whenever the conditions for an exact match do not apply and the following condition is met:
  - if $synsim(c_j, k_j) = s \geq t_a$, with s as the match score;

where *lowercase*(n), *lemma*(n) and *synsim*(n,m) are functions applied on concepts which respectively return a lowercase representation of the given concept, return its lemma form via a dictionary-based approach, and compute the syntactical similarity between the given concepts via the normalized Jaccard distance.

Once the algorithm is through, the third and final phase, or Phase 2, takes place, by considering the three lists of matches returned from Phase 1 (the latter two ordered by decreasing match score) as follows:

- M_e, list of exact matches; for each of those, a RT relationship "isExactMatchOf" is produced between the concepts matched;
- M_p, list of partial matches; for each of those, a RT relationship "isPartialMatchOf" is produced between the concepts matched;
- M_a, list of approximate matches; for each of those, a RT relationship "isApproximateMatchOf" is produced between the concepts matched.

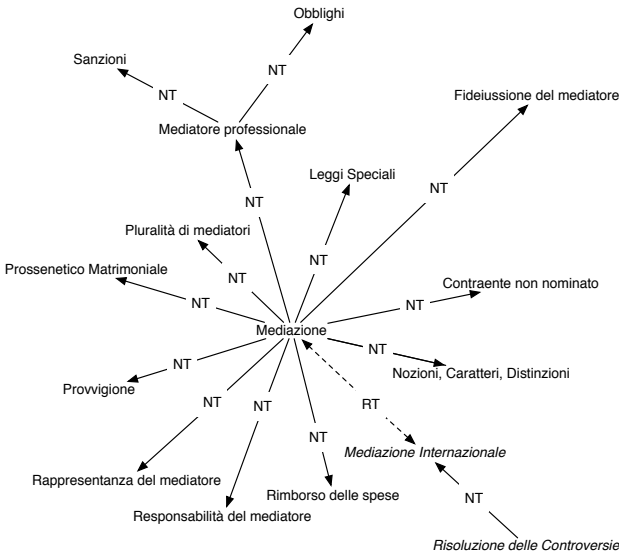

Figure 2. Excerpt of the legal ontology focused on mediation.

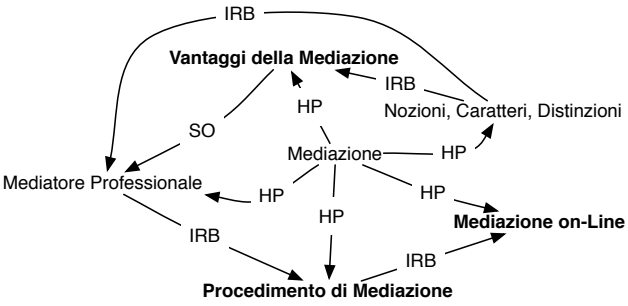

Figure 3. Excerpt of the legal ontology focused on educational relationships.

Figure 2 shows an excerpt of the integrated ontology focused on the concept of mediation, where the concept in italics comes from *EuroVoc*, whereas all the other concepts are from *ItalGiure*; the connection between the italicized concept and the corresponding ItalGiure concept is generically expressed in the figure as a *RT* relationship with dashed lines. Educational information has been manually added to the model for a small subset of concepts for which training modules were available.

Figure 3 shows a subset of concepts covering the theme of civil mediation from the educational point of view. Concepts in bold have been added to the model for educational purposes. Relationships from the educational set have also been added.

The relationships state that, in order to learn the concept of [mediation] ("mediazione"), it is necessary to learn the five sub-concepts [notions, types, distinctions] ("nozioni, caratteri, distinzioni"), [mediation process] ("procedimento di mediazione"), [advantages of mediation] ("vantaggi della mediazione"), [professional mediator] ("mediatore professionale") and [online mediation] ("mediazione online"). Besides, to approach both the [advantages of mediation] and the [professional mediator] concept, it is necessary to learn the [notion, types, distinctions] concept first. To learn about the [mediation process] it is needed to have understood the concept of [professional mediator]. To learn about [online mediation], it is needed to have understood the [process of mediation] concept. To favor learning, it is also desirable to teach [advantages of mediation] before the [professional mediator] concept.

C. Definition of additional metadata

As seen at the beginning of this section, *atomic legal resources* that are used to build KPs may be of different kinds. While the integrated ontology derived from the *ItalGiure* and the *EuroVoc* repository provides enough legal principles, precepts and sentences to be used for informative purposes, additional resources (i.e. learning objects) have been added for educational purposes and included in an external *educational repository*.

Resources contained in both repositories are indexed through metadata that, for each resource, provides additional information as summarized in Table I. In addition to that, metadata are used to link the available resources with ontological concepts the resources might deal with (see the *Concepts* field).

TABLE I. METADATA SCHEMA

Metadata Field	**Feasible Values**
Identifier	Unique identifier of the resource within the *legal ontology* or the *educational* repositories
Resource Type	A value in the set {legal principle, precept, sentence, learning object}
Description	Free text
Didactic method	Only for learning objects: a value in the set {presentation, test, exercise, storytelling, video}
Interactivity level	Only for learning objects: a value in the set {low, medium, high}
Concepts	A list of concepts of the legal ontology the resource refers to.

To be effectively used in conjunction with external resources and repositories, the adopted schema for learning resources has been mapped on the *IEEE LOM (Learning Object Metadata)*[5] standard.

D. Ontology enrichment via common-sense knowledge

This final step is aimed at calculating the set of weighted terms $T(c)$ characterizing each ontological concept, by exploiting common-sense knowledge held in a corresponding knowledge repository: for our purposes, *Wikipedia* has been used. In other words, for each ontological concept c, this step calculates:

$$T(c) = \{(t_1^c, w_1^c), (t_2^c, w_2^c), \ldots, (t_n^c, w_n^c)\}$$

where each t_i is a link to a *Wikipedia* topic (i.e. an article page) while each w_i measures the relevance of the topic with respect to the concept c.

To do that, the description $D(c)$ of each ontological concept is analyzed and the most relevant *n-grams* (sequences of n words) are selected based on the *keyphraseness* defined as the probability for an *n-gram* to be a *Wikipedia* link [17]:

$$Key(n\text{-}gram) = \frac{Link(n\text{-}gram)}{Count(n\text{-}gram)}$$

where $Link(n\text{-}gram)$ is the number of topics in which the *n-gram* appears as a link in *Wikipedia* while $Count(n\text{-}gram)$ is the number of topics in which the *n-gram* appears.

All *n-grams* with a *keyphraseness* over a threshold are considered *candidate topic referrals* for the concept c. Unfortunately, a *n-gram* may refer to different *Wikipedia* topics. In order to select the right topic for each concept, a disambiguation process is needed. This is done by relying on two different measures: *relatedness* and *commonness*.

For each pair of topics t_x and t_y, the sets X and Y of all hyperlinks that appear in the text of the topics are identified and their overlap $X \cap Y$ is calculated. Let N be the total number of *Wikipedia* topics, the *relatedness* between t_x and t_y is defined as follows:

$$Rel(t_x, t_y) = 1 - \frac{max(log|X|, log|Y|) - log|X \cap Y|}{N - min(log|X|, log|Y|)}$$

Furthermore, the commonness of a topic t_x with respect to a given *n-gram* is defined as:

$$Com(t_x|n\text{-}gram) = \frac{Link(n\text{-}gram|t_x)}{Link(n\text{-}gram)}$$

where $Link(n\text{-}gram|t_x)$ is the number of topics in which the *n-gram* appears as a link to the topic t_x.

Let Ctx be the set of *context topics* (i.e. topics associated with candidate referrals that do not require disambiguation), it is possible to associate a score with each topic t associated with an ambiguous candidate referral *n-gram* in this way:

$$Score(t|n\text{-}gram) = \frac{\sum_{t' \epsilon Ctx} Rel(t, t')}{|Ctx|} \times Com(t|n\text{-}gram)$$

For each candidate topic referral, a new pair (t^c, w^c) is thus included in $T(c)$ where t^c is the referred Wikipedia topic with the maximum score according to the previous equation while w^c corresponds to the value of $Key(n\text{-}gram)$.

IV. KNOWLEDGE PATH GENERATION

A *Knowledge Path* (KP) is composed of a training path and a set of additional information resources. By relying upon the legal ontology described in Section III, a KP is generated at run-time from the description of a legal case given as input. The KP generation process can be summarized in the three steps described below and outlined in Figure 4.

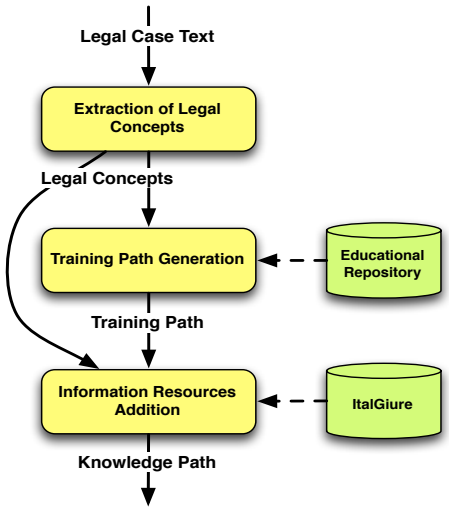

Figure 4. Knowledge path generation process schema.

a) *Extraction of legal concepts*. Firstly, the system tries to detect relevant legal concepts from the legal ontology that may be connected to the current legal case. A weight is associated to each detected concept, and expresses the latter's relevance in the given case.
b) *Training path generation*. Secondly, by taking advantage of the extracted legal concepts from the legal ontology and the available learning objects, the system generates a *training path* aimed at providing the basis for understanding legal issues related to the input case.
c) *Addition of information resources*. Finally, the system enriches the *training path* by adding relevant legal information like legal principles, precepts and sentences.

The next subsections provide additional details about all of these steps.

A. Extraction of legal concepts

This step is aimed at the detection of legal concepts from a given case expressed in a natural language text s. The output is a set $C(s)$ of pairs (c_i, μ_i) where c_i is a concept from C and μ_i specifies the relevance of c_i within the

[5] http://ltsc.ieee.org/wg12/

given text. The first operation is the extraction of a list of weighted terms from *s*:

$$T(s) = \{(t_1^s, w_1^s), (t_2^s, w_2^s), \dots, (t_n^s, w_n^s)\}$$

where, as seen is Section III.D, each t_i is a link to a *Wikipedia* topic, whereas each w_i measures the relevance of the topic with respect to the given text. The process to calculate $T(s)$ is the same used to calculate $T(c)$, but taking the given text as input rather than the concept description.

According to [18], to detect concepts in *s*, for each $c \in C$, a matching is performed between $T(s)$ and $T(c)$ through a combination of the standard measures of *precision* (P) and *recall* (R) [19] with the following equations:

$$P(s,c) = \frac{\sum_{t_j^c = t_k^s} (w_j^c * w_k^s)}{\sum_{t_j^c = t_k^s} (w_j^c)}$$

$$R(s,c) = \frac{\sum_{t_j^c = t_k^s} (w_j^c * w_k^s)}{\sum_{t_j^c = t_k^s} (w_j^s)}$$

The relevance score μ between *c* and *s* is then calculated with the following equation:

$$\mu = \frac{1}{|T(s)|} R(s,c) + \left(1 - \frac{1}{|T(s)|}\right) P(s,c)$$

Concepts from the legal ontology that have a relevance score over a given threshold (heuristically set to 0.5) are added to $C(s)$ together with the relevance score itself.

B. Training path generation

The generation of the *training path* is done starting from the set of concepts $C(s)$ extracted from the input text. The process, according to [20], is split into three subsequent steps.

The *first step* is aimed at building, from the ontology *O*, the simplified graph *O'*(*C*, *HP'*, *IRB'*, *SO'*) where *HP'* is the inverse relationship of *HP*, *IRB'* and *SO'* are initially set to *IRB* and *SO* but they are modified by applying the following rule: each arc $ab \in IRB' \cup SO'$ is substituted with arcs *ac* for all $c \in C$ such that there exist a path from *c* to *b* on the arcs from *HP'*. Figure 5a shows the graph *O'* obtained from the ontology reported in Figure 3.

The *second step* is aimed at building the graph *O''*(*C'*, *R*) where *C'* is the subset of *C* including all concept that must be taught according to $C(s)$ i.e. *C'* is composed by all nodes of *O'* from which there is a ordered path in $HP' \cup IRB'$ to concepts in *C(s)*. *R* is initially set to $HP' \cup IRB' \cup SO'$ but all arcs referring to concepts external to *C'* are removed. Figure 5b shows the graph *O''* obtained from *O'* starting from the target concept "*Mediazione on-Line*".

The *third step* finds a linear order between nodes of *O''* by using depth-first search so by visiting the graph nodes along a path *P* as deep as possible. Then it deletes from the obtained path all non-atomic concepts, i.e. all concept *a* so that $ab \in$ HP for some *b*. This ensures that only leaf concepts (with respect to the *HP* relationship) will be part of path *P*. Figure 5c shows the path *P* obtained from the graph *O''*.

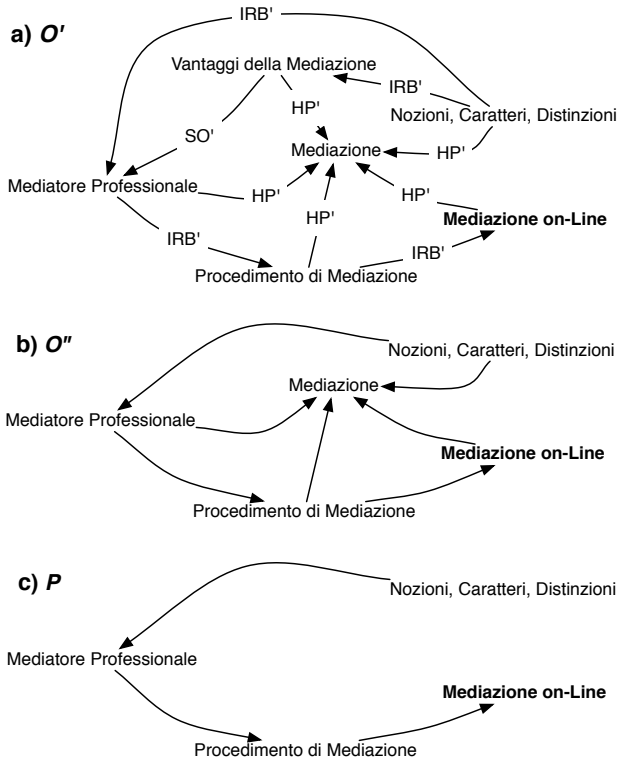

Figure 5. Example of the training path generation process.

By relying on the metadata described in Table I the system then finds feasible learning objects covering concepts in *P*. The obtained sequence is the *training path*.

C. Addition of information resources

In this step, once the *training path* has been generated, additional informative resources are added to it in order to obtain the KP. To do that, first of all, the set *D* of legal resources connected to each concept c_i of $C(s)$ is retrieved from the legal ontology by using the metadata shown in Table I.

Let $\{c_1^d, c_2^d, \dots, c_m^d\}$ be the list of all concepts connected with a resource $d \in D$ according to the metadata, a matching rate $rank_d$ is calculated for each retrieved resource based on the relevance μ_i associated with each concept c_i of $C(s)$ through the following equation:

$$rank_d = \sum_{c_i^d = c_j} \mu_j$$

Legal resources with a rank value greater then a threshold (heuristically set to 0.5) are added to the *training path*. The obtained structure is the resulting KP.

V. SYSTEM AND EXPERIMENTATION

Figure 6 shows a screenshot of the developed system. The user can write the legal case text in the uppermost box. In the box below, concepts extracted from text are displayed together with the relevance score μ. In this case, a legal case about a car accident is provided as input and the concepts of *insurance*, *provisional driving license* and *circulation of vehicles* are detected by the system with the higher score.

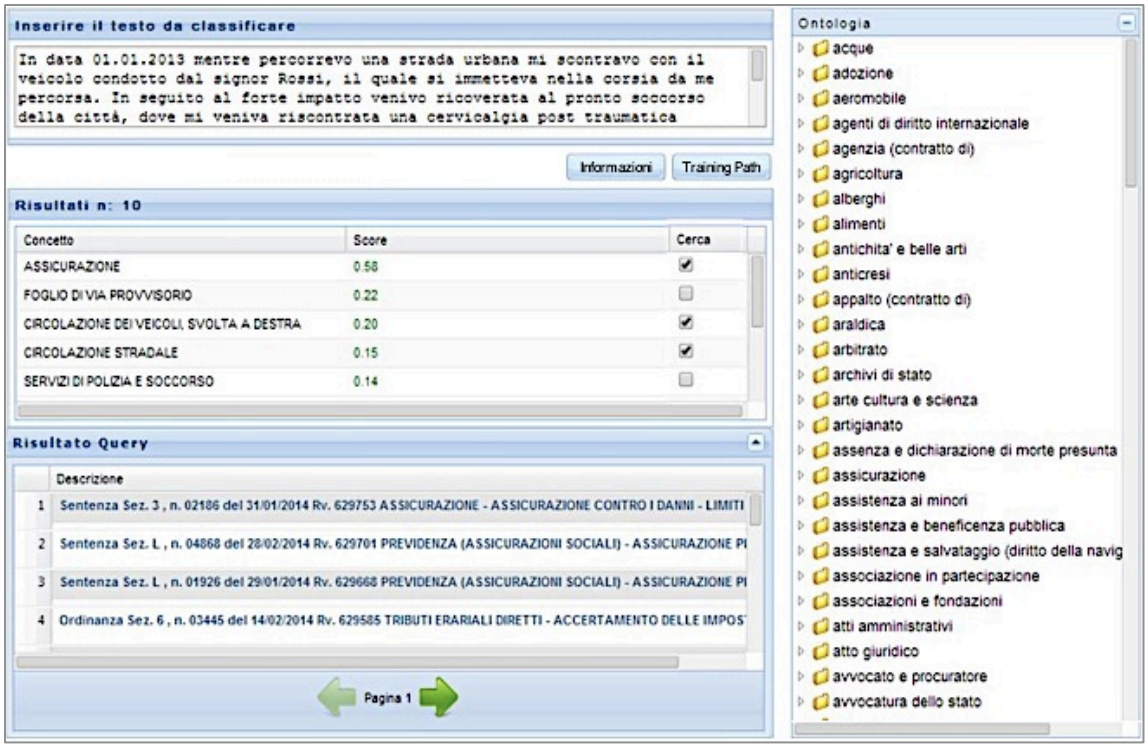

Figure 6. Snapshot of the prototype system.

By clicking on each concept, the legal ontology on the right box opens and displays the position of the selected concept within the NT hierarchy. Moreover, in the lower-most box, connected legal resources are displayed. By clicking on every resource it is possible to display it from the ItalGiure repository (Figure 7 shows a sample precept connected with the three detected concepts).

By clicking on the training path button, the user can access another page where she can follow a beginner course generated on the fly to cover principles connected with extracted concepts (on-line mediation and civil liability of the motorist in the specific case). Within this page (see Figure 8) he can chose any learning object from the sequence on the left and read the selected resource in the main panel.

A first experimentation has been performed to evaluate the effectiveness of the search process connected with the *information resources addition* phase described in IV.C. To evaluate algorithm performances, 50 queries generates from different input text have been analyzed.

A preliminary experimentation phase has been carried out in order to evaluate the effectiveness of the search process connected with the *Addition of information resources* described in Section IV.C. For this purpose, 50 queries generated from different input texts have been analyzed.

The quality of the approach is assessed in terms of precision and recall measures, considering the analysis through micro-average of the individual precision-recall curves [19]. Let $Q = \{Q_1, Q_2, \ldots, Q_n\}$ be a set of queries, and D all the relevant resources for the given set of queries Q. For each query Q_i, $\lambda = 20$ steps are considered, up to its maximum recall value, and measure the number of relevant documents retrieved at each step λ.

The micro-averaging of recall and precision (at the generic step λ), is defined as follows:

$$Rec_\lambda = \sum_{Q_i} \frac{|R_{Q_i} \cap B_{\lambda,Q_i}|}{|R|} \quad Prec_\lambda = \sum_{Q_i} \frac{|R_{Q_i} \cap B_{\lambda,Q_i}|}{|B_\lambda|}$$

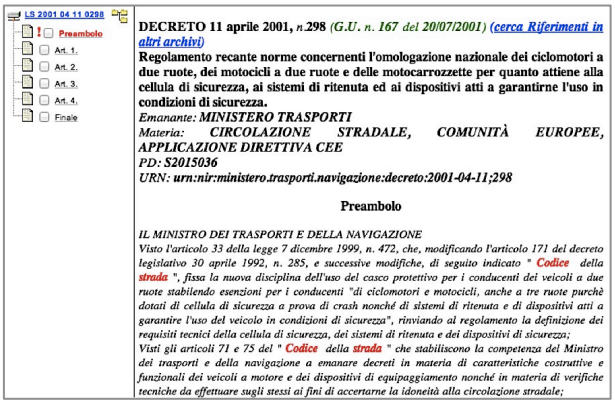

Figure 7. Example of a legal information resource

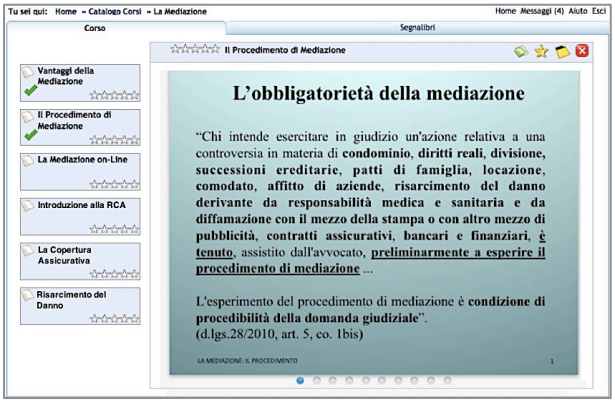

Figure 8. Snapshot of the training path delivery page.

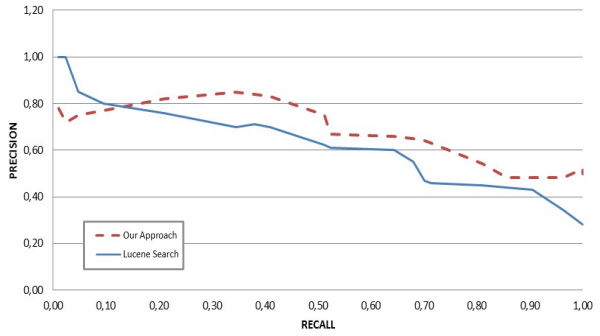

Figure 9. Micro-averaging precision/recall

where R_{Qi} is the set of relevant resources for a given query Q_i, B_λ the set of retrieved resources at the step λ and $B_{\lambda,Qi}$ is the set of all relevant resources, retrieved at the step λ, for the query Q_i.

Figure 9 shows the tendency of the micro-average of recall/precision curve evaluated on the collection set, and compares the approach used with a well-known keyword-based search engine called *Lucene*[6].

It is important to stress out that the effectiveness and accuracy of the semantic search is strongly dependent on the quality of the underlying legal ontology, along with the enrichment mechanism. The latter, as testified by the results, improves its effectiveness as the length of the input text query increases, since the more information is available, the better the tool is able to correctly understand its context and come up with meaningful concepts.

In this regard, the prototype of the system earlier described has been enhanced by including advanced functionalities related to the management of legal ontologies. Specifically, the system features a panel for managing different versions of the considered legal ontologies (*ItalGiure* and *EuroVoc* at the moment, as well as their integrated form), by which it is possible to perform both coarse-grained and fine-grained modifications and updates to the latter (Figure 10).

In fact, the system allows for editing and deleting existing concepts or adding new ones; it can execute the enrichment process either for an entire version of an ontology or for specific concepts; it enables the user to freely choose one of the managed ontologies to be used for the building of the knowledge path earlier described, allowing for a cross-experimentation among different ontologies or different versions of them.

Furthermore, the ontology integration algorithm detailed in Section III.A is available to be applied at runtime on any pair of ontologies in the system, in order to produce, for instance, an updated version of the integrated ontology including only matches that might be relevant to the user (e.g. only the exact matches, or a given subset of the partial and approximate matches, since the user is given the chance of manually reviewing each of them before confirming their inclusion).

Performance in terms of execution time of the process for the generation of KPs is almost instantaneous for texts of the considered length (up to 2000 characters). Additional tests, for longer legal cases, are yet to be performed and splitting mechanisms might be needed to maintain the execution time at acceptable levels.

[6] http://lucene.apache.org

Further tests are underway for other components of the system, as well as for the system as a whole from a user's perspective. In this regard, two additional experimentation phases are currently being carried out: the first involves a number of legal experts, in order to assess the quality and accuracy of the results produced by the automatic processes of the system (including the knowledge path generated and the matched concepts of the ontologies); the second is instead aimed at general users and is meant to check the effectiveness, usability and the overall user-friendliness of the features provided by the system. The real-world data collected in these phases will be used to further improve the system and finalize it for its release.

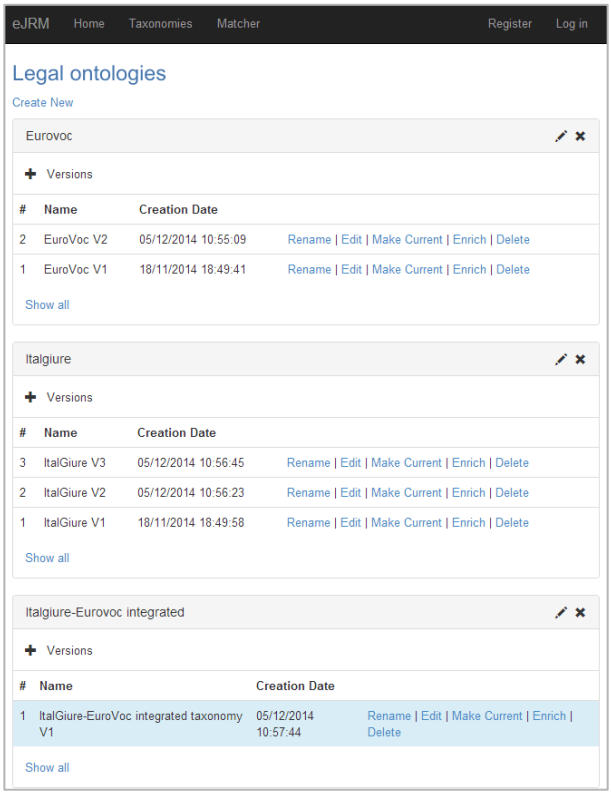

Figure 10. Screenshot showing the panel for managing legal ontologies and their versions.

VI. CONCLUSION

This paper describes a methodology for the automatic building of training, or "knowledge" paths, from natural language descriptions of legal cases, by using a legal ontology built for such a purpose and an enrichment process relying on common-sense knowledge.

The legal ontology has been defined and built by integrating and enriching two major legal sources, that is to say EuroVoc and ItalGiure, and the process used to produce it has been thoroughly detailed. In order to fill the *language gap* between legal language and that of common citizens, the enrichment process has been proposed to add common-sense meaning, taken from the *Wikipedia* knowledge base, to each ontological concept.

Results from a preliminary experimentation phase have been reported, and some additional functionalities of the system implementing the methodology have been presented, including the management of different versions for the legal ontology and the execution of a matching algorithm

between two input legal sources. Additional results related to the extraction of legal concepts can be found in [18], whereas a shorter version of the whole discussion, as mentioned earlier, is in [3].

It must be underlined that the component of the system, responsible for managing, integrating, enriching and exploiting the legal can handle any well-formed ontologies and apply upon them its integration, versioning and enrichment mechanisms, regardless of their original contexts. As such, this component can be in principle be used to tackle other, different domains that may require functionalities for ontology management, editing, integration, enrichment and versioning. In this regard, the authors are specifically exploring the possibility of extending the results obtained on the legal domain to areas like Human Resource Management and Human Capital Management.

Integration of this system with the other components planned for the eJRM project is currently underway, and further experimentation and validation in this regard has been scheduled and will be presented in future works.

REFERENCES

[1] G. Arosio, G. Bagnara, N. Capuano, E. Fersini, D. Toti, "Ontology-driven Data Acquisition: Intelligent Support to Legal ODR Systems", in Proc. Int. Conf. on Legal Knowledge and Information Systems (JURIX 2013), IOS Press, pp. 25-28, 2013.

[2] N. Capuano, C. De Maio, A. Gaeta, G.R. Mangione, S. Salerno, E. Fratesi, "A Storytelling Learning Model for Legal Education", in Proc. Int. Conf. on e-Learning (e-Learning 2014), ACM, 2014.

[3] N. Capuano, S. Salerno, C. De Maio. "A Knowledge Based System for Guidance and Training on Legal Concepts", Proceedings of the 6th International Conference on Intelligent Networking and Collaborative Systems (INCOS 2014), IEEE Computer Society, pp. 498-503, 2014. http://dx.doi.org/10.1109/INCoS.2014.32

[4] R.W. van Kralingen, "A conceptual frame-based ontology for the law". Proc. Int. Workshop on Legal Ontologies (LEGONT 97), pp.15-22, 1997.

[5] A. Valente, J. Breuker, "A functional ontology of law", in Towards a Global Expert System in Law, Cedam Publishers, 1994.

[6] J. Hage, B. Verheij, "The law as a dynamic interconnected system of states of affairs: A legal top ontology", in Int. J. of Human-Computer Studies, vol. 51, n. 6, pp. 1043-1077, 1999. http://dx.doi.org/10.1006/ijhc.1999.0297

[7] A.S. Ferrer, J.M.M. Rivero, A.M. García, "Improvements in Recall and Precision in Wolters Kluwer Spain Legal Search Engine", in Computable Models of the Law, Springer, pp. 130-145, 2008. http://dx.doi.org/10.1007/978-3-540-85569-9_9

[8] R. Rubino, A. Rotolo, G. Sartor "An OWL ontology of fundamental legal concepts" in Proc. Int. Conf. on Legal Knowledge and Information Systems (JURIX 2006), IOS Press, pp. 101-110, 2006.

[9] S. van de Ven, R. Hoekstra, J. Breuker, L. Wortel, A. El-Ali, "Judging Amy: Automated legal assessment using OWL 2", in Proc. of OWL: Experiences and Directions (OWLED 2008), 2008.

[10] G. Ajani, G. Boella, L. Lesmo, M. Martin, A. Mazzei, D.P. Radicioni, P. Rossi, "Legal Taxonomy Syllabus version 2.0", Proc. 3rd ICAIL Workshop on Legal Ontologies and Articial Intelligence Techniques, CEUR, 2009.

[11] M. Poblet, P. Casanovas, J.M.L. Cobo, "Online Dispute Resolution for the Next Web Decade: The Ontomedia Approach", in Proc. 10th Int. Conf. on Knowledge Management and Knowledge Technologies, 2010.

[12] M.A. Biasiotti, M.T. Sagri, D. Tiscornia "LOIS: Building a Multilingual Wordnet for the Legal Domain", in Proc. 4th Int. EGOV Conference 2005, Aug. 22-26, 2005.

[13] E. Francesconi, D. Tiscornia, "Building Semantic Resources for Legislative Drafting: The DALOS Project", in Computable Models of the Law, LNCS, Springer, vol. 4884, pp 56-70, 2008.

[14] P. Casanovas, G. Sartor, N. Casellas, R. Rubino (Eds.), "Computable Models of the Law", Springer, 2008. http://dx.doi.org/10.1007/978-3-540-85569-9

[15] N. Capuano, M. Gaeta, P. Ritrovato, S. Salerno, "How to Integrate Technology Enhanced Learning with Business Process Management", J. of Knowledge Management, Emerald, vol. 12 n. 6, pp. 56-71, 2008. http://dx.doi.org/10.1108/13673270810913621

[16] N. Capuano, G.R. Mangione, A. Pierri, S. Salerno, "Personalization and Contextualization of Learning Experiences basing on Semantics", Int. J. of Emerging Technologies in Learning, vol. 9, n. 7, pp. 5-14, 2014. http://dx.doi.org/10.3991/ijet.v9i7.3666

[17] D. Milne, I. Witten, "An open-source toolkit for mining Wikipedia" in Proc. New Zealand Computer Science Research Student Conf., NZCSRSC, 2011.

[18] N. Capuano, C. De Maio, S. Salerno, D. Toti, "A Methodology based on Commonsense Knowledge and Ontologies for the Automatic Classification of Legal Cases", Proc. Int. Conf. on Web Intelligence, Mining and Semantics (WIMS 2014), 2014. http://dx.doi.org/10.1145/2611040.2611048

[19] C.J. van Rijsbergen "Information retrieval" (2nd ed.). University of Glasgow: Dept. of Computer Science, 1979.

[20] N. Capuano, M. Gaeta, A. Marengo, S. Miranda, F. Orciuoli, P. Ritrovato, "LIA: an Intelligent Advisor for e-Learning", Interactive Learning Environments, T&F, vol. 17, n. 3, pp. 221-239, 2009.

AUTHORS

Nicola Capuano is scientific officer at the University of Salerno. His main research interest is artificial intelligence and, among its applications, intelligent tutoring systems and knowledge representation. He works as a project manager and research consultant within several research and development projects. He is author of several scientific papers. He is scientific referee and member of editorial boards for International journals and conferences (e-mail: ncapuano@unisa.it)

Andrea Longhi is an ICT Engineer and Computer-Science Researcher. After obtaining the Master's Degree in Computer Engineering on advanced Semantic Web technologies, he has been selected as one of the best Italian graduates for the Campus Mentis 2010. Specialized in DTA technologies, he is now a Research Fellow at the University of Salerno, where he worked in various projects such as TITAN, HSEPGEST and eJRM (e-mail: alonghi@unisa.it).

Saverio Salerno is Full Professor of Mathematics at the Faculty of Engineering of the University of Salerno. He is author of several papers regarding Mathematical Analysis, Operational Research, Simulation, Learning, Knowledge, Computer Science. He has been the Italian Delegate for the IST programme of the EU FP5; he has been expert of the European Commission for Education and Training and member of the Advisory Committee of the Italian Ministry of Education, Universities and Research (e-mail: salerno@unisa.it).

Daniele Toti holds a Ph.D. in Computer Science and Engineering and is currently a Senior Research Scientist at the University of Salerno. His research activity mainly revolves around information extraction and knowledge discovery, as well as ontology building, matching and integration. Relevant projects he participated in include Aristotele, SIRET and eJRM. He is author of several papers in international journals and conference proceedings, holds a CPE from the University of Cambridge and is an Oracle and Sun Certified Professional (e-mail: dtoti@unisa.it).

The work presented in this paper has been partially supported by the eJRM (electronic Justice Relationship Management) project, co-funded by the Italian Ministry of Instruction and Research (ref. PON 01-01286). This article is an extended and modified version of a paper presented at the International Workshop on Adaptive Learning via Interactive, Collaborative and Emotional approaches (ALICE 2014), held on September 10-12, 2014, Salerno, Italy. Submitted 08 April 2015. Published as resubmitted by the authors 05 May 2015.