These are your rights:
A Natural Language Processing Approach to Automated RDF Licenses Generation

Elena Cabrio[1,2], Alessio Palmero Aprosio[3], and Serena Villata[1]

[1] INRIA Sophia Antipolis, France
firstname.lastname@inria.fr
[2] EURECOM, France
[3] Machine Linking Srl, Italy
alessio@machinelinking.com

Abstract. In the latest years, the Web has seen an increasing interest in legal issues, concerning the use and re-use of online published material. In particular, several open issues affect the terms and conditions under which the data published on the Web is released to the users, and the users rights over such data. Though the number of licensed material on the Web is considerably increasing, the problem of generating machine readable licenses information is still unsolved. In this paper, we propose to adopt Natural Language Processing techniques to extract in an automated way the rights and conditions granted by a license, and we return the license in a machine readable format using RDF and adopting two well known vocabularies to model licenses. Experiments over a set of widely adopted licenses show the feasibility of the proposed approach.

Keywords: #eswc2014Cabrio

1 Introduction

The material published on the Web is usually associated to its terms of use and re-use, which provide the legal permissions and requirements the user has to comply with when dealing with such material. In the Web of Data, the majority of the published datasets are associated to specific licenses: as it has been shown in [9, 17], about 75% of all Linked Data datasets listed in the CKAN archive[4] (Comprehensive Knowledge Archive Network) is associated to a license. Specifying the terms of re-use of the data is particularly important to foster the use and re-use of the data itself, as underlined in [11]. However, apart from the problem of specifying the license under which a certain dataset is released, other problems arise in the actual licenses and copyright specification in the Web of Data scenario, and in the Web in general. In particular, despite the Linked Data principles [2], only few datasets are associated to the machine readable version of the adopted license. As discussed by Rodriguez-Doncel et al. [17], specific

[4] http://datahub.io/

licensing terms are still referenced in natural language (NL) text, and there is the need to provide tools for supporting users in producing rights expressions in a machine readable format, such that more datasets could be easily associated to licenses. The lack of machine readable licenses specifications affects also the development and adoption of frameworks dealing with the licensing terms in an automated way, like for instance the licenses compatibility and composition framework proposed by Governatori et al. [9].

In this paper, we answer the following research question:

– How to support the creation of machine readable licensing information, starting from the natural language specification of the licenses?

The first point to be addressed consists in deciding the language to adopt to specify the licenses in a machine readable format. We choose to rely on the RDF language[5], since it is a standard model for data interchange on the Web. Moreover, it is the language adopted by the Creative Commons Rights Expression Language (CC REL)[6], that explains how license information may be described in a machine readable format using RDF. We aim at supporting both human users and automated systems to generate, starting from the natural language specification of the licenses, their RDF counterpart. Our scenario is as follows. On the one side, we have a human user publishing a dataset on the Web of Data; she wants to release its dataset for instance under the Open Government License[7] and, to be compliant with the Linked Data principles [2], she wants to specify in RDF such license. This means she has to know the possible vocabularies able to express licensing terms, and she has to go through the license text "translating" natural language terms into RDF. On the other side, an automated tool, like those presented in [19, 9], retrieves a number of datasets on the Web of Data, and it needs the licensing information about such data. The problem is that each dataset only provides, e.g., in its VoID description[8], only the link to the natural language text of the license. In order to retrieve processable licensing information, it has to crawl the natural language text and automatically build its RDF description. Therefore, our research question breaks down into the following sub-questions: *i)* Which vocabularies have to be adopted to express licenses in RDF?, and *ii)* How to develop an automated framework to support the generation of RDF licenses specifications from their natural language texts?

First, we analyze existing vocabularies to represent licensing information, and we choose two of them, namely the Creative Commons Rights Expression Language Ontology[9], and the Open Digital Rights Language (ODRL) Ontology[10]. In particular, we adopt the former to represent in RDF Creative Commons (CC) licenses, and the latter to model all other licenses, given its broader scope. Our

[5] `http://www.w3.org/2011/rdf-wg/wiki/Main_Page`
[6] `http://wiki.creativecommons.org/CC_REL`
[7] `http://www.nationalarchives.gov.uk/doc/open-government-licence/`
[8] `http://www.w3.org/TR/void/`
[9] `http://creativecommons.org/ns`
[10] `http://www.w3.org/ns/odrl/2/`

RDF-based representation of licenses represents the specific rights (i.e., permissions, prohibitions and duties) granted by the licenses in the NL text.

Second, we adopt Natural Language Processing (NLP) techniques to develop an automated online framework called *NLL2RDF* (*Natural Language License to RDF*) able to "translate" natural language licenses specifications into their RDF definition using either the ODRL or the CC REL vocabulary. More precisely, NLL2RDF relies on machine learning techniques: the task is treated as a classification problem in supervised learning, and the adopted learning algorithm is Support Vector Machines (SVM) [4]. The algorithm is then trained over a set of manually annotated licenses.

The proposed approach is a first attempt to provide an automated framework able to output the RDF representation of the natural language description of a license. NLL2RDF is intended to support the diffusion of RDF-based licensing information attached to the datasets published on the Web of Data. Moreover, our approach is not limited to the Web of Data scenario, but it can be used to provide machine readable representation of licensing information not only for datasets but also for documents or software products, e.g., the Apache License[11]. Note that, given the complexity of the task, the current version of NLL2RDF provides an RDF representation of licenses considering their basic deontic components only, i.e., we model *permissions*, *prohibitions*, and *duties* only, and we do not consider at the present stage further constraints expressed by the licenses, e.g., about time, payment information, and sub-licensing. The automated treatment of such information is left as future work.

The remainder of the paper is as follows. Section 2 describes the vocabularies, and details the architecture of the proposed RDF licenses generation framework. Section 3 presents the experimental setting, and the evaluation of NLL2RDF. Section 4 compares the proposed approach with the related work in the literature.

2 Licenses: from terms and conditions to triples

In this section, we first motivate our choice of the ODRL and CC REL vocabularies, and we then describe the classes and properties we adopt from such vocabularies (Section 2.1). Finally, we present the proposed framework to translate NL licenses into their RDF representation (Section 2.2).

2.1 CC REL and ODRL vocabularies

Several vocabularies have been proposed in the last years to model licensing information. In particular, the following interconnected vocabularies provide high level descriptions of licenses, with a particular attention to the Web of Data scenario: LiMO[12], L4LOD[13], and ODRS[14]. More complex licenses information can

[11] `http://www.apache.org/licenses/LICENSE-2.0`
[12] `http://data.opendataday.it/LiMo`
[13] `http://ns.inria.fr/l4lod/`
[14] `http://schema.theodi.org/odrs/`

be defined with one of the digital Rights Expression Languages like ODRL[15] or MPEG-21, a machine-readable language that allows to declare rights and permissions using the terms as defined in the Rights Data Dictionary.[16] These vocabularies, ODRL in particular, have not been specifically conceived for the Web of Data scenario, but they intend to provide flexible mechanisms to support transparent and innovative use of digital content in publishing, distribution and consumption of digital media across all sectors. So far only the CC REL [1], the standard recommended by CC for the machine-readable expression of licensing terms, has been used by the Linked Data community.

We choose ODRL and CC REL vocabularies for our purposes. The reasons are the following: *i)* CC REL is the vocabulary to be used for all CC licenses, and it is the mostly adopted vocabulary in the Linked Data community for licenses specification; and *ii)* ODRL allows the specification of fine grained licensing terms both for data (thus satisfying the Web of Data scenario), and for all other digital media, allowing a broader application of NLL2RDF.

CC REL specifies for each `cc:License` a set of `cc:Permissions` (an action that may or may not be allowed), `cc:Requirements` (an action that may or may not be requested to the user), and `cc:Prohibitions` (something the user is asked not to do). The vocabulary specifies the following permissions (`cc:Reproduction`, `cc:Distribution`, `cc:DerivativeWork`, `cc:Sharing`), requirements (`cc:Notice`, `cc:Attribution`, `cc:ShareAlike`, `cc:SourceCode`, `cc:Copyleft`, `cc:LesserCopyleft`), and prohibitions (`cc:CommercialUse`, and `cc:HighIncomeNationUse`). For more details on the CC REL vocabulary, see [1]. Let us consider a license, like CC Attribution-NonCommercial License[17], where permissions are Reproduction, Distribution and Derivative Works, requirements are Notice and Attribution, and Commercial Use is prohibited. The license is represented in RDF (Turtle syntax[18]) using the CC REL vocabulary as follows:[19]

```
:licCC-BY-NC a cc:License;
         cc:legalcode <http://creativecommons.org/licenses/by-nc/4.0/>;
         cc:permits cc:Reproduction;
         cc:permits cc:Distribution;
         cc:permits cc:DerivativeWorks;
         cc:requires cc:Notice;
         cc:requires cc:Attribution;
         cc:prohibits cc:CommercialUse.
```

ODRL specifies, instead, different kinds of Policies (i.e., Agreement, Offer, Privacy, Request, Set and Ticket). In NLL2RDF we adopt `Set`, a policy expression that consists in entities from the complete model. Permissions, prohibitions and duties (i.e., the requirements specified in CC REL) are specified in terms of

[15] `http://www.w3.org/community/odrl/`
[16] `http://iso21000-6.net/`
[17] `http://creativecommons.org/licenses/by-nc/4.0/`
[18] `http://www.w3.org/TeamSubmission/turtle/`
[19] Prefixes are omitted for clarity reasons.

an `action`. For instance, we may have the action of attributing an `asset` (anything which can be subject to a policy), i.e., `odrl: action odrl: attribute`. For more details about the ODRL vocabulary, refer to the ODRL Community group.[20] The following example shows a Set policy expression, stating that the licensed asset is the target of the permission to reproduce, distribute, derive, the duty to attribute and attach the policy and, the prohibition to commercialize. It expresses the same rights as the CC license reported above.

```
:licCC-BY-NC a odrl:Set;
        odrl:permission [
            a odrl:Permission;
            odrl:action odrl:reprodice;
            odrl:action odrl:distribute;
            odrl:action odrl:derive
        ] ;
        odrl:prohibition [
            a odrl:Prohibition;
            odrl:action odrl:commercialize
        ] ;
        odrl:duty [
            a odrl:Duty;
            odrl:action odrl:attribute;
            odrl:action odrl:attachPolicy
       ] .
```

The NLL2RDF framework will adopt CC REL vocabulary to specify CC licenses, and the ODRL vocabulary to specify all other licenses.

2.2 The NLL2RDF framework

After choosing the vocabularies to be used to express licenses in RDF, we can now describe our framework for RDF-based licenses specifications automatically extracted from natural language texts. The architecture of NLL2RDF is visualized in Figure 1. NLL2RDF can be accessed both by humans through the web interface[21] and by automated tools through the API of the system.

NLL2RDF input is the natural language definition of the licensing terms to be "translated" into RDF. NLL2RDF access such NL text T and applies some preprocessing steps: tokenization, lemmatization, part-of-speech tagging. After that, a classification step is performed, using kernel methods. We first embed the input data in a suitable feature space, and then use a linear algorithm to discover nonlinear patterns in the input space. Typically, the mapping is performed implicitly by a so-called *kernel function*.

Formally, the kernel is a function $k : X \times X \rightarrow \mathbb{R}$ that takes as input two data objects (e.g., vectors, texts, parse trees) and outputs a real number characterizing

[20] `http://w3.org/ns/odrl/2/`
[21] A demo of NLL2RDF is available at `www.airpedia.org/NLL2RDF`

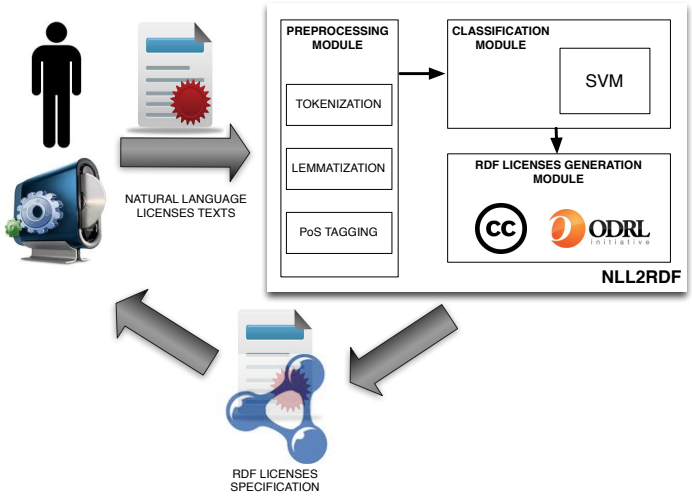

Fig. 1. The architecture of NLL2RDF.

their similarity, with the property that the function is symmetric and positive semi-definite. That is, for all x_1, $x_2 \in X$, it satisfies

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle,$$

where ϕ is an explicit mapping from X to an (inner product) feature space $\mathcal{F}$. For our task, we work at sentence level, meaning that each sentence is considered as a vector. We define and combine two different kernel functions that calculate the pairwise similarity between sentences (*bag-of-words* and *verb*). Many classifiers can be used with kernels, we use Support Vector Machine. In particular, we used libSVM 3.12[22], a freely available tool.

The simplest method to calculate the similarity between two sentences is to compute the inner product of their vector representation in the vector space model (VSM). Formally, we define a space of dimensionality N in which each dimension is associated with one feature, and the sentence s is represented by a row vector

$$\phi_j(s) = (w(f_1, s), w(f_2, s), \ldots, w(f_N, s)), \tag{1}$$

where the function $w(f_k, s)$ records whether a particular feature f_k is active in the sentence s. Using this representation, we define the *bag-of-features kernel* between sentences as

$$K_F(s_1, s_2) = \langle \phi_j(s_1), \phi_j(s_2) \rangle, \tag{2}$$

***Bag-of-words* kernel** The *bag-of-words kernel* (K_W) is defined as in Equation (2) where the function $w(f_k, s)$ in Equation (1) is the standard *term frequency-inverse document frequency* (tf $\times$ idf) of the word f_k in the sentence s.

[22] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

***Verb* kernel** The *verb kernel* (K_V) is defined as in Equation (2) where the f_k in Equation (1) are elements of the set including the union of all the verb tokens (part-of-speech tags starting with "V") and the same tokens preceded by the token "not" in the sentence. Each verb token t is associated to two different features, depending on whether it is preceded by "not". In particular, if s contains that token, the corresponding feature is activated (that is $w(f_k, s) = 1$).

Composite Kernel Having defined the two basic kernels representing different characteristics of entity descriptions, we finally define the composite kernel as

$$K_{\text{COMBO}}(s_1, s_2) = K_W(s_1, s_2) + K_V(s_1, s_2) \tag{3}$$

The individual kernels are normalized. This plays an important role in allowing us to integrate information from heterogeneous feature spaces. It follows directly from the explicit construction of the feature space and from closure properties of kernels that the composite kernel is a valid kernel.

NLL2RDF returns to the querying agent the RDF description of the licensing terms provided in natural language. Note that NLL2RDF does not provide any triple about the licensed work/asset. This means that, in case the generated RDF license has to be used to license a specific asset `asset841`, then a triple concerning the connection between the license and the asset has to be added by the agent (human or automated) who uses NLL2RDF.

3 Experimental setting

To experiment our framework NLL2RDF, we selected a set of licenses (i.e. all the licenses adopted to certify data in the Linked Data cloud, plus additional software and online published material licenses) to create our reference dataset (described in Section 3.1). We then run NLL2RDF to generate the machine readable version of these licenses. More details on the experiments, and a discussion of the results we have obtained are reported in Section 3.2.

3.1 Dataset creation

In order to evaluate NLL2RDF, as a first step we selected a set of licenses to build our reference dataset. More specifically, our reference dataset is composed by 37 licenses, comprising all the licenses adopted to certify data in the Linked Data cloud (as all the Creative Commons licenses[23]), software licenses (as Mozilla Public License[24] and Microsoft License[25]), and additional licenses for other material on the Web (as the UK Open Government license, and the New Free Documentation License[26]).

[23] `http://creativecommons.org/licenses/`
[24] `http://www.mozilla.org/MPL/2.0/`
[25] `http://referencesource.microsoft.com/referencesourcelicensing.aspx`
[26] `http://www.gnu.org/copyleft/fdl.html`

As a second step, we manually "translated" the textual version of each license into RDF, adopting the vocabularies described in Section 2.1 (i.e. CC REL for Creative Commons and ODRL for all the other licenses). Given for instance a textual fragment of the ODC Open Database License (ODbL)[27]:

You are free: To Share: To copy, distribute and use the database. To Create: To produce works from the database. To Adapt: To modify, transform and build upon the database. As long as you: Attribute: You must attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database. Share-Alike: If you publicly use any adapted version of this database, or works produced from an adapted database, you must also offer that adapted database under the ODbL. [...]

we manually built the machine readable version of the license as follows:

```
@prefix odrl:  http://www.w3.org/ns/odrl/2/.
@prefix : http://example/licenses.

:licODBL a odrl:Set;
         odrl:permission [
           a odrl:Permission;
           odrl:action odrl:derive;
           odrl:action odrl:share
       ] ;
       odrl:duty [
           a odrl:Duty;
           odrl:action odrl:attribute;
           odrl:action odrl:shareAlike
       ] .
```

We use this machine readable version of the licenses as a goldstandard, i.e., to be compared with NLL2RDF's output in order to evaluate its ability in generating a correct RDF from the licenses texts.

As a third step in the creation of the reference dataset, we annotated in the textual version of the license the sentences containing the lexicalization of the ontological relations (i.e., the sentences whose meaning correspond to the ontological relations), to train our system. For instance, in the example of the ODbL license above, we annotated the sentence *You are free: To Share the database* with the ODRL relation `odrl:Permission` and the value `odrl:share`; the sentence *You are free: To produce works from the database* with the ODRL relation `odrl:Permission` and the value `odrl:derive`; the sentence *As long as you: Attribute: You must attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL* with the ODRL relation `odrl:Duty` and the value `odrl:attribute`; and the sentence *As long*

[27] `http://opendatacommons.org/licenses/odbl/summary/`

as you: Share-Alike: If you publicly use any adapted version of this database, or works produced from an adapted database, you must also offer that adapted database under the ODbL with the ODRL relation `odrl:Duty` and the value `odrl:shareAlike`.

The same annotation task has been carried out on Creative Common licenses adopting CC REL ontology. For instance, given a textual fragment of the Attribution 4.0 International (CC BY 4.0) license[28]:

You are free to: Share - copy and redistribute the material in any medium or format. Adapt - remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms. Under the following terms: Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. No additional restrictions You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

we manually built the machine readable version of the license as follows:[29]

```
@prefix cc: http://creativecommons.org/ns.
@prefix : http://example/licenses.

:licCC-BY a cc:License;
             cc:legalcode <http://creativecommons.org/
                                              licenses/by/4.0/legalcode>;
             cc:requires cc:Notice;
             cc:requires cc:Attribution;
             cc:permits cc:Reproduction;
             cc:permits cc:Distribution;
             cc:permits cc:DerivativeWorks.
```

We then annotate in the textual version of the license the sentences whose meaning correspond to the ontological relations: the sentence *You are free to: copy the material in any medium or format* with the CC REL relation `cc:permits` and the value `cc:Reproduction`; the sentence *You are free to: redistribute the material in any medium or format* with the CC REL relation `cc:permits` and the value `cc:Distribution`; the sentence *You are free to: remix, transform, and build upon the material for any purpose* with the CC REL relation `cc:permits` and the value `cc:DerivativeWorks`; the sentence *You must provide a link to the license* with the CC REL relation `cc:requires` and the value `cc:Notice`; the sentence *You must give appropriate credit* with the CC REL relation `cc:requires` and the value `cc:Attribution`.

[28] `http://creativecommons.org/licenses/by/4.0/`

[29] CC licenses are also available in XML/RDF format on the CC website. CC-BY in particular is available at `http://creativecommons.org/licenses/by/4.0/rdf`

For the dataset annotation we adopted the CONLL IOB format[30], usually used in the NLP community for Natural Language Learning shared tasks. We first tokenized the sentences using Stanford Parser [18], and we then added two columns, the first one for the annotation of the relation, and the second one for the value, as follows[31]:

```
#id-004
1 You         PRP    B-PERMISSION      DERIVE
2 are         VBP    I-PERMISSION
3 free        JJ     I-PERMISSION
4 :           :      O
[...]                O
5 To          TO     I-PERMISSION
6 produce     VB     I-PERMISSION
6 works       VBZ    I-PERMISSION
7 from        IN     I-PERMISSION
8 the         DT     I-PERMISSION
15 database   NN     I-PERMISSION
16 .          .      O
```

The dataset has been annotated and independently verified by two annotators, with a complete agreement on the annotations (as introduced before, at this stage NLL2RDF considers licenses' basic deontic components only, for which human agreement is complete on almost all of them).

3.2 Evaluation

In our experiments, we use a linear SVM classifier for each possible relation-value present in all the licenses. In addition, we mapped the CC REL vocabulary labels on the ODRL labels, to increase the number of examples to train and test NLL2RDF (we apply then the mapping in the reverse order to generate the correct RDF for CC REL licenses). Only the couples relation-values with more than 5 occurrences in the data are reported.[32] Table 3.2 describes the NLL2RDF performances in the relation assignment over the licenses included in our dataset.

[30] In this scheme, each token is tagged with one of three special chunk tags, I (inside), O (outside), or B (begin). A token is tagged as B if it marks the beginning of a chunk. Subsequent tokens within the chunk are tagged I. All other tokens are tagged O. The B and I tags are suffixed with the chunk type according to our annotation task, e.g. B-PERMISSION, I-PERMISSION. Of course, it is not necessary to specify a chunk type for tokens that appear outside a chunk, so these are just labeled O.

[31] The annotated dataset is available at `www.airpedia.org/NLL2RDF/dataset-licenses`. Each couple relation-value has been annotated in a separate file, contained in a folder with the license name.

[32] The following couples relation-values with less than 5 occurrences in the data are: `Permission:read` (1 occurrence), `Permission:commercialize` (3), `Permission:share` (4), `Duty:attachSource` (1), `Prohibition:distribute` (3), and `Prohibition:modify` (1).

Given a sentence, we test it against every classifiers, so that we can intercept those sentences containing more than one relation (see Section 3.1 for an example). Performances are calculated using the n-fold cross-validation ($n = 3$). The annotated data set is randomly split into 3 parts containing the same number of examples (1/3 of the total, around 560 sentences). A single subset is retained as test set, while the remaining 2 subsets are used as training data. The process is executed 3 times, each time with a different subset used for validation, giving 3 different pairs of precision/recall values. These values are then averaged to obtain the final results.

Table 1. Performances of NLL2RDF on the correct assignment of each triple.

relation-value	**# occurr.**	**precision**	**recall**	**f-measure**
`Permission:distribute`	28	0.74	0.59	0.65
`Permission:derive`	15	0.66	0.51	0.56
`Permission:reproduce`	14	0.55	0.51	0.46
`Permission:modify`	13	0.66	0.2	0.3
`Permission:copy`	11	0.77	0.22	0.34
`Permission:sell`	6	0.83	0.38	0.53
`Duty:shareAlike`	17	0.72	0.3	0.36
`Duty:attachPolicy`	16	0.76	0.63	0.68
`Duty:attribute`	15	1	0.66	0.78
`Prohibition:commercialize`	7	1	0.33	0.49

NLL2RDF reaches quite interesting results for some relation-value assignment, mainly for the ones with a high number of occurrences in the training data (e.g. `Permission:distribute`, `Duty:attachPolicy`, `Duty:attribute`). For some other relations, SVM performances are far from being optimal, due to *i)* the sparsity of some relations in the data (i.e. for some couples relation-value only few examples are present in the data, e.g. `Prohibition:commercialize`), *ii)* the fact that the lexicalizations of relations such as `Permission:modify` involve a lot of language variability, each one not supported by a sufficient number of occurrences in the text (e.g. *you are free to modify*; *assure everyone the effective freedom [...] with modification*), and *iii)* very similar surface forms can refer to different relations-values (for instance for `Duty:shareAlike` and `Duty:attachPolicy`, we have the textual representations *Redistributions must reproduce the above copyright notice* for the former, and *Redistributions must retain the copyright notice* for the latter). We are aware that the current version of NLL2RDF is not yet fully reliable, but at the present stage our system is not yet intended to completely substitute users: it is intended as a tool to support them in specifying the machine readable version of licensing information. As a short improvement, we are planning to collect and annotate other licenses, to increase our training dataset in size and variability. Moreover, since certain structures in licenses

appear over and over, we are envisaging to add manually written rules to capture recurrent patterns. In general, more efforts would also be required from the community, to encourage data providers to publish machine readable licenses, semi-automatically and with the support of the NLL2RDF system.

4 Related Work

Heath and Bizer [11] underline that "the absence of clarity for data consumers about the terms under which they can reuse a particular dataset, and the absence of common guidelines for data licensing, are likely to hinder use and reuse of data". Therefore, all Linked Data on the Web should include explicit licenses, or waiver statements, as discussed by [13], who propose the Open Data Commons licenses that try to fully license any rights that cover databases and data.

Beside the vocabularies mentioned in Section 2.1, other few vocabularies have been proposed in the literature to model, to different extent, licensing information. The Waiver vocabulary[33], for instance, defines properties to use when describing waivers of rights over data and content, where a waiver is defined as a voluntary relinquishment or surrender of some known right or privilege. As underlined by [9, 17], licenses are usually connected to the data through the VoID description. In particular, the Dublin Core vocabulary[34] is usually adopted to associate licenses to resources through the property `dc:license`, and the class `dc:LicenseDocument` provides the legal document giving official permission to do something with the resource. Two further vocabularies which may be adopted to define the licensing terms associated to the data on the Web are the Description of a Project vocabulary (DOAP)[35], and the Ontology Metadata vocabulary (OMV)[36]. More precisely, DOAP specifies a property `doap:license` referring to the URI of an RDF description of the license the software is distributed under; OMV defines the property `omv:hasLicense`, which provides the underlying license model, and a class `omv:LicenseModel`, which describes the usage conditions of an ontology. The attachment of additional information like rights or licenses to RDF triplets may be done also by adopting named graphs [3]. Carroll et al. [3] introduce them to allow publishers to communicate assertional intent and to sign their assertions. Moreover, the W3C Provenance WG [10] defines the kind of information to be used to form assessments about data quality, reliability or trustworthiness.

The different licenses, e.g., Creative Commons, Open Data Commons, have common features, but also differ from each others. The requirement to mention the author (attribution) seems to be one of the best shared features. Most legal frameworks allow commercial use: that is, they make it possible for re-users to sell public data without transforming or enriching them. The Web NDL Authority

[33] `http://vocab.org/waive/terms/`
[34] `http://dublincore.org/documents/dcmi-terms/`
[35] `http://usefulinc.com/ns/doap`
[36] `http://omv2.sourceforge.net/index.html`

license[37] is an exception, and prohibits reuse as is of its data for commercial purposes: a further individual examination by the licensor is necessary. For a further discussion about rights declaration in Linked Data, see [17].

In the Web scenario, a number of works address the problem of representing and/or reasoning over licensing information. Iannella [38] presents the Open Digital Rights Language (ODRL) for expressing rights information over content, and Gangadharan et al. [5] further extend ODRL developing the ODRL-S language to implement the clauses of service licensing. Gangadharan et al. [6] address the issue of service license composition and compatibility analysis basing on ODRL-S. They specify a matchmaking algorithm which verifies whether two service licenses are compatible. If so, the services can be composed and the framework determines the license of the composite service. Truong et al. [19] address the issue of analyzing data contracts, based again on ODRL-S. Contract analysis leads to the definition of a contract composition where first the comparable contractual terms from the different data contracts are retrieved, and second an evaluation of the new contractual terms for the data mash-up is addressed. Krotzsch and Speiser [12] present a semantic framework for evaluating ShareAlike recursive statements. In particular, they develop a general policy modelling language, then instantiated with OWL DL and Datalog, for supporting self-referential policies as expressed by CC. Finally, Gordon [8] presents a legal prototype for analyzing open source licenses compatibility using the Carneades argumentation system. All these works assume licensing information to be expressed in some kind of machine readable format or formal syntax. Given that licenses are always expressed first in natural language, these frameworks could rely on NLL2RDF to have a first machine readable version of the license in RDF. Then, the "translation" from RDF to the specific formal syntax they need to reason over licensing terms has to be performed. However, this step is usually less demanding than a direct translation form NL to a specific syntax, given the high complexity and variability of natural language texts.

Closer to the general purpose of our work of supporting users in defining machine readable descriptions of licenses, Nadah et al. [14] propose to assist licensors' work by providing a generic way to instantiate licenses, independent from specific formats. Then they translate such license into more specific terms compliant with the specific standards used by distribution systems, i.e., ODRL and MPEG Rights Data Dictionaries. They do not address the problem of providing an automated tool to move from NL licenses to their RDF representation, but they propose a model to move from a license description through a particular ontology to the description of the same license using another ontology.

Rodriguez-Doncel et al. [16, 15] discuss licenses patterns for Linked Data. They first analyze and discuss six rights expression languages, abstracting their commonalities and outlining their underlying pattern. Second, they propose the License Linked Data Resources pattern which provides a solution to describe existing licenses and rights expressions both for open and not open scenarios.

[37] `http://iss.ndl.go.jp/ndla/use/`
[38] `http://odrl.net/1.1/ODRL-11.pdf`

Even if our final goal is different from theirs, the LLDR pattern is useful for an overall structured representation of the different rights expression languages.

5 Conclusions

In this paper, we presented NLL2RDF, an automated framework to support RDF-based licenses specifications starting from natural language texts. The goal of NLL2RDF is to provide both human users, and automated systems with a support to generate machine readable representations of licensing terms. We adopt the CC REL and the ODRL vocabularies to specify the licenses in RDF. Our framework relies on NLP techniques to generate in an automated way such RDF based licenses descriptions. In particular, the framework exploits SVM to classify the couples relation-value present in the licenses, and then the RDF version of the license is generated filling a pre-defined RDF template. In order to train the system, two annotators independently marked up a set of 37 licenses, selected among the set of widely adopted licenses in the Web of Data in particular, and in the Web in general. The experimental evaluation shows the feasibility of the proposed framework and fosters to pursue with this research direction. Both the dataset and the system, as web service, are available online.

NLL2RDF provides a first step towards the automatic analysis of natural language licenses texts to return their machine readable description. However, several open challenges still remain to be addressed. For instance, user evaluation is the first step of future works, even if we have already started to gather feedback about the systems' results from legal experts in the Web area. Second, we will extend the dataset to train our system by adding other licenses in order to improve the performances of our system, particularly with respect to those deontic components which do not appear frequently nowadays in the dataset. We are planning to collect and annotate other licenses, to increase our training dataset (so that to capture enough language variability to improve the system robustness). Third, we will improve the precision of RDF licenses description. As we previously motivated, at the present time we model licenses using only the basic deontic components they express without taking into account any further constraint or exception stated in the NL license text. Moreover, we plan to couple machine learning algorithms with pattern-based approaches for information extraction (following [7]). Finally, the system can be extended to a multilingual scenario (as far as a NLP tool to process the language at issue is available), to provide machine readable versions of licenses published by national institutions, or licenses published in different languages.

Acknowledgements

The work of Elena Cabrio was funded by the French Government (National Research Agency, ANR) through the "Investments for the Future" Program reference # ANR-11-LABX-0031-01.

References

1. H. Abelson, B. Adida, M. Linksvayer, and N. Yergler. ccREL: The creative commons rights expression language. Technical report, Creative Commons, 2008.
2. C. Bizer, T. Heath, and T. Berners-lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 2009.
3. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs. *J. Web Sem.*, 3(4):247–267, 2005.
4. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2010.
5. G. R. Gangadharan, V. D'Andrea, R. Iannella, and M. Weiss. Odrl service licensing profile (ODRL-S). In *Proceedings of Virtual Goods*, 2007.
6. G. R. Gangadharan, M. Weiss, V. D'Andrea, and R. Iannella. Service license composition and compatibility analysis. In *Proceedings of ICSOC, LNCS 4749*, pages 257–269. Springer, 2007.
7. D. Gerber and A.-C. N. Ngomo. Extracting multilingual natural-language patterns for rdf predicates. In *EKAW*, pages 87–96, 2012.
8. T. F. Gordon. Analyzing open source license compatibility issues with Carneades. In *Proceedings of ICAIL*, pages 51–55. ACM, 2011.
9. G. Governatori, A. Rotolo, S. Villata, and F. Gandon. One license to compose them all - a deontic logic approach to data licensing on the web of data. In *Proceedings of ISWC, LNCS 8218*, pages 151–166. Springer, 2013.
10. P. T. Groth, Y. Gil, J. Cheney, and S. Miles. Requirements for provenance on the web. *IJDC*, 7(1):39–56, 2012.
11. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.
12. M. Krötzsch and S. Speiser. ShareAlike Your Data: Self-referential Usage Policies for the Semantic Web. In *Proceedings of ISWC, LNCS 7031*, pages 354–369. Springer, 2011.
13. P. Miller, R. Styles, and T. Heath. Open data commons, a license for open data. In *Proceedings of LDOW*, 2008.
14. N. Nadah, M. D. de Rosnay, and B. Bachimont. Licensing digital content with a generic ontology: escaping from the jungle of rights expression languages. In *Proceedings of ICAIL*, pages 65–69. ACM, 2007.
15. V. Rodriguez-Doncel, M. S. Figueroa, A. Gomez-Perez, and M. P. Villalon. License linked data resources pattern. In *Proc. of the 4th International Workshop on Ontology Patterns*, 2013.
16. V. Rodriguez-Doncel, M. S. Figueroa, A. Gomez-Perez, and M. P. Villalon. Licensing patterns for linked data. In *Proc. of the 4th International Workshop on Ontology Patterns*, 2013.
17. V. Rodríguez-Doncel, A. Gómez-Pérez, and N. Mihindukulasooriya. Rights declaration in linked data. In O. Hartig, J. Sequeda, A. Hogan, and T. Matsutsuka, editors, *COLD*, volume 1034 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
18. R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In *ACL (1)*, pages 455–465, 2013.
19. H. L. Truong, G. R. Gangadharan, M. Comerio, S. Dustdar, and F. D. Paoli. On analyzing and developing data contracts in cloud-based data marketplaces. In *Proceedings of APSCC, IEEE*, pages 174–181, 2011.