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

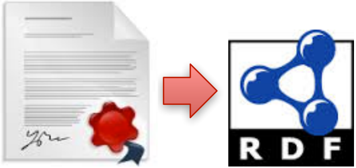
Elena Cabrio, Alessio Palmero Aprosio, Serena Villata

INRIA Sophia Antipolis, France
Machine Linking, Italy

From natural language licenses to machine readable ones

Research Question

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



Research Question

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

1. Which vocabularies to use to express licenses in RDF?
2. How to develop an automated framework to support the generation of RDF licenses from natural language texts?

From Terms and Conditions to Triples

Vocabularies for expressing licenses:

- Creative Commons (CC)
- Open Digital Rights Language (ODRL)
- LiMO
- L4LOD
- ODRS

From Terms and Conditions to Triples

Vocabularies for expressing licenses:

- **Creative Commons (CC)**
- **Open Digital Rights Language (ODRL)**
- LiMO
- L4LOD
- ODRS

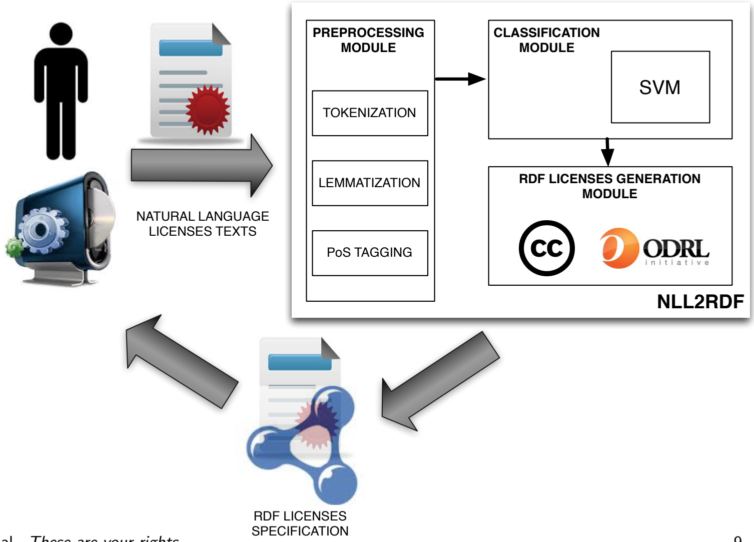
BY-NC license: example using CC

```
: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.
```

BY-NC license: example using ODRL

```
: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
       ] .
```

NLL2RDF architecture



NLL2RDF framework

1. Preprocessing steps:
   - tokenization, lemmatization, part-of-speech tagging
2. Classification step (sentence level):
   - Bag-of-words kernel and Verb kernel for pairwise similarity
     - **Bag-of-words**: standard term frequency-inverse document frequency of word f_k in the sentence s
     - **Verb**: union of all verb tokens (PoS) and same tokens preceded by "not" in the sentence
   - Composite Kernel: $K_{\mathrm{COMBO}}(s_1, s_2) = K_W(s_1, s_2) + K_V(s_1, s_2)$

NLL2RDF online tool:
`http://www.airpedia.org/nll2rdf-tool/`

NLL2RDF framework

1. Preprocessing steps:
   - tokenization, lemmatization, part-of-speech tagging
2. Classification step (sentence level):
   - Bag-of-words kernel and Verb kernel for pairwise similarity
     - **Bag-of-words**: standard term frequency-inverse document frequency of word f_k in the sentence s
     - **Verb**: union of all verb tokens (PoS) and same tokens preceded by "not" in the sentence
   - Composite Kernel: $K_{\mathrm{COMBO}}(s_1, s_2) = K_W(s_1, s_2) + K_V(s_1, s_2)$

NLL2RDF online tool:
`http://www.airpedia.org/nll2rdf-tool/`

Dataset creation (37 licenses from LOD cloud datasets)

Manual dataset creation

Example: Open Database License (ODbL)
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. [...]

Dataset creation (37 licenses from LOD cloud datasets)

```
@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
       ] .
```

Dataset creation
(37 licenses from LOD cloud datasets)

Dataset annotation

```
#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
```

Evaluation

Performances of NLL2RDF on the correct assignment of each triple (3-fold cross validation, 1/3 of the total: around 560 sentences)

relation-value	#occurr.	P	R	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

Evaluation: error analysis

- sparsity of some relations in the data (i.e. only few examples are present in the data, e.g. `Prohibition:commercialize`),
- lexicalizations of relations e.g. `Permission:modify` involve a lot of language variability, without sufficient number of occurrences in the text (e.g. *you are free to modify; assure everyone the effective freedom [...] with modification*),
- very similar surface forms can refer to different relations-values (e.g. 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)

Evaluation: error analysis

- sparsity of some relations in the data (i.e. only few examples are present in the data, e.g. `Prohibition:commercialize`),
- lexicalizations of relations e.g. `Permission:modify` involve a lot of language variability, without sufficient number of occurrences in the text (e.g. *you are free to modify; assure everyone the effective freedom [...] with modification*),
- very similar surface forms can refer to different relations-values (e.g. 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)

Evaluation: error analysis

- sparsity of some relations in the data (i.e. only few examples are present in the data, e.g. `Prohibition:commercialize`),
- lexicalizations of relations e.g. `Permission:modify` involve a lot of language variability, without sufficient number of occurrences in the text (e.g. *you are free to modify; assure everyone the effective freedom [...] with modification*),
- very similar surface forms can refer to different relations-values (e.g. 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)

Conclusions

NLL2RDF

- automated framework to **support RDF-based licenses specifications** starting from natural language texts

Future perspectives

- User evaluation of NLL2RDF tool
- Extend dataset to improve the performances
- Improve precision of RDF licenses descriptions
- Couple ML algorithms with pattern-based approaches

Conclusions

NLL2RDF

- automated framework to **support RDF-based licenses specifications** starting from natural language texts

Future perspectives

- User evaluation of NLL2RDF tool
- Extend dataset to improve the performances
- Improve precision of RDF licenses descriptions
- Couple ML algorithms with pattern-based approaches

Thanks for your attention!