Singling out Legal Knowledge from World Knowledge. An NLP–based approach

Francesca Bonin[*][◊], Felice Dell'Orletta[°], Giulia Venturi[°] and Simonetta Montemagni[°]

[*] *Università di Pisa, Dipartimento di Informatica – Pisa*

[◊] *Language Interaction and Computation Lab, University of Trento*

[°] *Istituto di Linguistica Computazionale "Antonio Zampolli", (ILC–CNR) – Pisa*

Abstract. Ontology learning in the legal domain rises the well-known problem of *epistemological promiscuity* between legal entities and regulated domain instances. In this paper, we propose a new term extraction approach specifically aimed at tackling such a problem through the acquisition of a term glossary where legal terms, expressing legal concepts, and domain terms, providing a description of the regulated world knowledge, are automatically singled out. The proposed approach has been tested with promising results on a corpus of Italian European legal texts regulating the environmental domain.

Keywords: Terminology Extraction, Natural Language Processing, Legal Ontology

1. Introduction

Scholars committed to modeling legal domain knowledge have widely acknowledged with the need for domain–specific knowledge organization, i.e. legal ontologies, where domain knowledge (*legal knowledge*) and knowledge of domains of interest to be regulated (referred to as *world knowledge*) are not mixed. However, as pointed out in Breuker et al. (2004), the indiscriminate mixture of the two types of knowledge is a common attitude in constructing legal ontologies. In particular, Breuker and colleagues speak of *epistemological promiscuity*, putting the emphasis on how this is a serious problem in core ontology development. They point out that many legal ontologies collapse together *epistemological and ontological perspectives*. Starting from the well-known assumption that "by its very nature, law deals with behaviour in the world", they discuss how domain independent concepts of law are tained with common–sense notions which refer to social activities. Interestingly, they claim that "the domain ontologies [they] developed in the various project contained almost ninety–nine percent terms that belonged to the category 'world knowledge', i.e. the world the legal domain is about". On the contrary, a core ontology should exclusively include "typical legal concepts, like norm, responsibility, person (agent), action, etc.". Moreover, the most serious consequence envisaged is that "ontologies mixed with epistemological frameworks have a far more limited re–use and may pose more interoperability problems than clean ontologies." In fact, the *level of gen-*

erality adopted in constructing a domain ontology is closely related to the reusability issue. According to the state of the art in ontology design criteria reported in Casellas (2008), several levels can be established ranging from the more abstract *top or upper–level* ontologies, which include general concepts not domain–specific, and *core* ontologies, which provide top–level domain–specific (i.e. legal) concepts, to *domain–specific* ontologies, which organize world knolwedge, providing a description of a specific domain of interest to be regulated.

Building on these emergent issues, Francesconi (2010) has recently proposed an approach to legal knowledge modeling based on the separation of legal and world knowledge and oriented to interoperability and reusability. According to the knowledge model suggested, two levels of conceptualization are envisaged: a Domain Independent Legal Knowledge (DILK) level, which provides a model for legal rules independently from the domain they apply to, and a Domain Knowledge (DK) level, which offers information and relationships among entities specific for a given regulated domain. This approach follows Biagioli (2009), who claims that a law simultaneously *describes* the occurring events and *regulates* them.

In this paper, we face the *epistemological promiscuity* problem at the level of the acquisition of terminological knowledge from legal texts. Instead of starting from ready–made epistemological and ontological concepts, which are defined *a priori* on the basis of domain–theoretical assumptions, we propose a term extraction approach overtly aimed at automatically discriminating legal terms from regulated–domain terms. The paper is organised as follows: in Section 2, we motivate the proposed approach by discussing the background literature. Section 3 presents our Terminology Extraction methodology, while the results of a term extraction experiment on a corpus of Italian European legal texts concerning the environmental domain are reported in Section 4. The evaluation of achieved results is discussed in Section 5.

2. Background and motivation

As widely acknowledged in the literature, terminology extraction is the first and most–established step in ontology learning from texts. To put it in Buitelaar et al. (2005) words, "terms are linguistic realizations of domain–specific concepts and are therefore central to further, more complex tasks". In this context, the peculiar challenge posed by legal texts consists in the fact that they simultaneously contain legal terms and regulated domain terms. When dealing with legal texts, the process of terminological acquisition thus needs to take into account two main issues: i) the extraction of terms corresponding to domain–relevant concepts, and ii) the identification of the specific domain they refer to (i.e. the regulated domain or the legal domain). We strongly believe that singling out legal terms, i.e. those which express *legal knowledge*, from terms of the specific domain being regulated, i.e. those which express *world knowledge*, represents a helpful starting point for any further construction of legal ontologies where *legal* and *world* knowledge is kept separate.

Differently from the community of legal ontology developers, to our knowledge the problem of *legal knowledge* mingled with *world knowledge* has been addressed only in a few cases within the terminology extraction literature, i.e. by Lame (2005) and Lenci et al. (2009). The NLP–based terminology extraction experiments from French Codes carried out in Lame (2005) and aimed at identifing legal ontology components resulted in the irrelevance of statistical indices (such as Term frequency or Tf, Inverse document frequency or idf, etc.) to single out legal terms from domain terms. In the analysis of results achieved with the T2K (*Text–to–Knowledge*) ontology learning system, Lenci et al. (2009) notice that, as expected from the peculiar nature of processed documents, the acquired term bank includes both legal and regulated–domain terms. Since the two classes of terms show quite different frequency distributions, several acquisition experiments were carried out by setting different thresholds: it turned out that terms belonging to the target domain regulated by law are always scarcely represented in the final result, due to their high rank (and low frequency) according to Zipf's law. Note however that, differently from Lame (2005), Lenci et al. (2009) main concern was not the classification of terms but rather the fact that both term types should be adequately represented in the final result.

To deal with the epistemological promiscuity problem and to overcome the aforementioned difficulties, we propose an approach simultaneously meant to acquire relevant terminology from legal texts and to discriminate between legal and regulated–domain terms. For this purpose, we follow the layered approach to terminology extraction described in Bonin et al. (2010), where, firstly, candidate terms are identified using state–of–the–art statistical measures and, secondly, a shortlist of well–formed and relevant candidate terms is reranked by applying a contrastive method. The goal of this paper is to show to what extent such a methodology is successful in acquiring from a corpus of Italian European legal texts concerning the environmental domain a term list where terms belonging to the legal domain (e.g. *disposizione nazionale* 'national provision', *disposizione di presente direttivo* 'provision of the present directive', etc.) and to the regulated environmental domain (e.g. *sostanza pericoloso* 'hazarous substance', *valore limite di emissione* 'emission limit value', etc.) are clearly singled out. Following Buitelaar et al. (2005), this can be the starting point to develop a domain ontology where concepts expressing *legal* and *world* knowledge are not mixed.

3. The term extraction approach

The term extraction method we followed, described in detail in Bonin et al. (2010), combines NLP techniques, linguistic and statistical filters. For our present purposes, we are interested both in one–word terms (single terms), e.g. *president*, as well as multi–word terms (complex terms), e.g. *president of republic*.

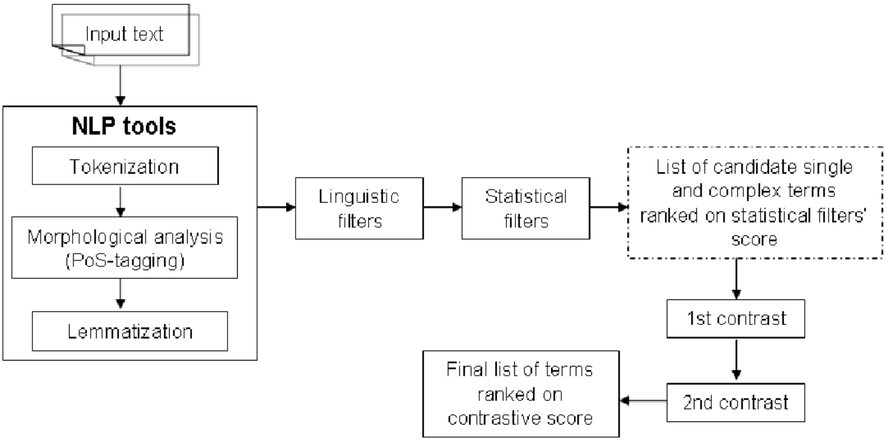
Input text
NLP tools
Tokenization
Morphological analysis (PoS-tagging)
Lemmatization
Linguistic filters
Statistical filters
List of candidate single and complex terms ranked on statistical filters' score
1st contrast
2nd contrast
Final list of terms ranked on contrastive score

Figure 1. Term Extraction Process

As shown in Figure 1, which illustrates the general extraction process, the input text is firstly tokenized, morphologically analyzed (i.e. PoS–tagged) and lemmatized passing through a pipeline of state–of–the–art NLP tools for the analysis of Italian texts. The PoS–tagged text, obtained with the tagger described in Dell'Orletta (2009), is searched for on the basis of linguistic filters aimed at identifying a) nouns, expressing candidate single terms and b) PoS patterns covering the main nominal modification types which express candidate complex terms. It is the case of morpho–syntactic templates such as noun + adjective (e.g. *decreto legislativo* 'legislative decree'), noun + preposition + noun (e.g. *decreto del presidente* lit. 'decree of the president'), etc.

At this stage, the candidate single terms are ranked on the basis of their frequency of occurrence in the input text, while the candidate complex terms are ranked on the score of a different statistical filter. For this purpose, the C-NC Value measure is used as described in Frantzi et al. (1999) and Vintar (2004). It is currently considered as the state–of–the–art method for terminology extraction and it is meant to assessing the likelihood for a term of being a well–formed and relevant multi–word term. Afterwards, the contrastive method is applied against the list of ranked candidate single and multi–word terms. As shown in Figure 1, where the intermediate output of the extraction process is displayed in a dotted box, the two top lists of candidate (single and multi-word) terms are contrasted firstly against the term list extracted from an open–domain corpus and secondly against a top list of terms acquired from a legal corpus differing at the level of the regulated domain. In both contrastive phases, the contrastive function (CSmw) newly introduced in Bonin et al. (2010) is used. The CSmw score is based on the arctangent function that tends to valorize less frequent data, and in fact reveled to be suitable for handling variation in low frequency events such as multi–words or regulated–domain terms. The first contrastive analysis stage (so–called "1st contrast") is meant to prune common words (if any) from the list of domain–relevant terms, while the second contrastive analysis stage (so–called "2nd contrast") allows obtaining a list of terms where regulated–domain and legal terminology is discriminated, being respectively at the top and at the bottom of the final term list.

4. Experiments and results

The term extraction methodology described above has been tested on a document corpus constituted by a collection of European legal texts of 394,088 word tokens concerning the environmental domain (hereafter referred to as "Environmental Corpus"). Following the extraction process illustrated in Section 3, for the first contrastive analysis stage we used as open–domain contrastive corpus the PAROLE Corpus (Marinelli et al., 2003), made up of about 3 million words and including Italian texts of different types (newspapers, books, etc.) testifying general language usage; for the second contrastive analysis stage, a corpus of 74,210 word tokens, containing European law texts on consumer protection (hereafter generically referred to as "Legal Corpus"), was used instead.

In the rest of the paper, we will focus on the extraction of multi–word terms. The reason for this choice is twofold: if on the one hand multi–word terms have been demonstrated to cover the vast majority of domain-specific terminology (85% according to Nakagawa et al. (2003)), on the other hand the proposed process of complex terms extraction highlights a number of novelties worth discussing further. As noted in Bonin et al. (2010), differently from previous studies which follow contrastive approaches, such as Basili et al. (2001), Penas et al. (2001) and Chung et al. (2004), we prefer basing complex term acquisition on their concrete occurrence in texts as unique elements separate from single terms. Althought this novelty is not the main focus of the present work, it is interesting to point out how this new method aims at extracting only those multi-words that are specifically relevant in the domain at hand. In fact, the relevant single term *principio* 'principle' is extracted.

However multi–words headed by this single term are not extracted, unless they are relevant themselves for the domain topic, differently from (Basili et al., 2001) where all multi–word terms, having a domain specific single head, are extracted, independently from their domain specificity; in other words, we will not extract terms such as *principio di precauzione* 'precautionary principle' and *principio fondamentale* 'fundamental principle' even if they occur in texts and share the same single head term (i.e. *principio* 'principle'). Instead we acquire complex terms such as *principio attivo* 'active ingredient' and *principio di sussidiarietà* 'principle of subsidiarity' that are relevant multi–word terms themselves.

In the extraction experiment we carried out, we started from the extraction of a list of well formed candidate multi-words, in line with the morpho–syntactic constraints we set. Then, we selected a top list[1] from the candidate term list ranked on score of the statistical filter, thus obtaining a shortlist of 600 either legal (e.g. *norma europea*, 'European norm'), environmental (e.g. *emissione di gas a effetto serra*, 'emission of greenhouse gases') or open–domain terms (e.g. *direttore generale*, 'director–general'). Afterwards, we firstly contrasted the top list of 600 multi–word terms against the top list extracted from the PAROLE Corpus, in order to reduce the noise deriving from highly frequent common words (e.g. *giorno successivo*, 'following day' or *anno precedente*, 'previous day'), obtaining a list mainly made of environmental and legal terms. Then, in order to distinguish environmental and legal terms, we contrasted a top list of 300 environmental–legal multi–word terms against the top list extracted from the Legal Corpus, obtaining a final list of 300 terms ranked on the contrastive score. In this final list, environmental terms were expected to be found at the top of the final list ranked according to the contrastive score, while the legal terms were expected at the bottom. Tables I and II report respectively the first and the last 10 multi–word terms of the final 300 multi–word term list we obtained after the second step of contrast. Interestingly enough, the top of the final list as reported in Table I contains environmental terms, represented by the first 10 multi–word terms extracted from the Environmental Corpus ranked according to their decreasing contrastive score. Table II shows the final part of the list, constituted by the legal terms (the 10 multi–word terms extracted from the Environmental Corpus ranked according to their increasing contrastive score). These results will be discussed in Section 5.

[1] Note that the thresholds we set up for this experiment were empirically defined and mainly meant to show to what extent the proposed approach was correctly working for what concerns the filtering of legal and environmental terms. It goes without saying that final thresholds should be defined by taking into account the size of the document collection as well as typology and reliability of expected results.

Table I. First 10 multi–word terms extracted from the Environmental Corpus ranked according to their decreasing contrastive score

Environmental terms	Contrastive ranking
sostanza pericoloso (*hazarous substance*)	1.57079625565
salute umano (*human health*)	1.57079624903
sviluppo sostenibile (*sustainable developement*)	1.57079623794
principio attivo (*active ingredient*)	1.57079622006
inquinamento atmosferico (*air pollution*)	1.57079621766
effetto serra (*greenhouse effect*)	1.57079621254
rifiuto pericoloso (*hazardous waste*)	1.57079620696
valore limite di emissione (*emission limit value*)	1.57079620548
corpo idrico (*water body*)	1.57079616937
cambiamento climatico (*climate change*)	1.57079615637

Table II. Last 10 multi–word terms extracted from the Environmental Corpus ranked according to their increasing contrastive score

Legal terms	Contrastive ranking
funzionamento di mercato interno (*functioning of national market*)	1.5707610035
disposizione nazionale (*national provision*)	1.57078159756
disposizione essenziale di diritto interno (*essential internal provision of national law*)	1.57078274091
testo di disposizione essenziale di diritto (*text of essential provision*)	1.57078274091
testo di disposizione (*text of provision*)	1.57078547573
diritto nazionale (*national law*)	1.57078699537
diritto interno (*national law*)	1.57078751378
livello di protezione (*level of protection*)	1.57078885837
disposizione di presente direttivo (*provision of the present directive*)	1.57079070201
norma nazionale (*national rule*)	1.57079084047

5. Evaluation

5.1. General evaluation criteria

The multi–word term list extracted from the Environmental Corpus has been evaluated in two different steps. First, it has been automatically compared against two different gold-standard resources selected for the environmental and legal domains. In particular, we used a) the thesaurus *EARTh (Environmental Applications Reference Thesaurus)*[2], containing 12,398 terms, as a reference resource for what concerns the environmental domain, and b) the *Dizionario giuridico* (Edizioni Simone) available online[3], including 1,800 terms, for the legal domain. Afterwards, those terms which have not been categorized as belonging to a specific domain during this automatic evaluation phase were manually validated by legal and environmental experts. These two different phases of evaluation were due to the fact that the considered reference resources have a good coverage of domain specific single terms (e.g. *disposizione*, 'provision', *valore* 'value', etc.), but they do not have a proper coverage of domain-specific complex terms (e.g. *disposizione essenziale del diritto*, 'law essential provision', *valore limite di emissione* 'emission limit value').

In order to evaluate how legal and environmental terms are distributed in the acquired 300–term list we further divided this list in 30–term groups. Interestingly, although the top list of 300 evaluated terms is quite small, it proved to be reliable in order to test to what extent the term extraction method we proposed can help to single out legal and regulated–domain terminology. However, we think that a future evaluation of a wider amount of extracted terms can provide more detailed insights into the distribution of the two types of terminology within a term list automatically acquired from legal corpora. Similarly, we can foresee an evaluation in terms of recall (calculated as the percentage of correctly acquired terms with respect to all terms in the gold standard lexicon): unfortunately, this type of evaluation poses so far a considerable problem due to the lack of a reference terminological resource aligned with respect to the acquisition corpus.

5.2. Discussion of results

The distribution of three different types of terms was evaluated. For each 30–term group of the final 300–term list we computed the amount of i) environmental terms, ii) legal terms, iii) terms which can refer to both domains, such as *politica ambientale*, 'environmental policy'. The remaining amount

[2] http://uta.iia.cnr.it/earth.htm#EARTh%202002

[3] http://www.simone.it/newdiz

of terms which were not categorized as belonging to types i), ii) or iii) are represented by errors.

Table III. Evaluation of the multi–word term list acquired from the Environmental Corpus

Group	Environmental	Legal	Environmental/Legal
0-30	16	5	3
30-60	17	3	3
60-90	12	2	3
90-120	8	9	2
120-150	14	7	1
150-180	9	12	2
180-210	15	3	3
210-240	11	12	1
240-270	9	14	1
270-300	0	22	1

As we can see in Table III which reports the distribution of the different term types within each single 30–term group, the adopted contrastive function is able to discriminate between environmental and legal terms. The first group contains 16 environmental terms against 5 legal terms; in the last group 22 legal terms and no environmental terms occur. This trend is pointed out in Figure 2, where the divergent lines show the different distributions of environmental and legal terms across the different 30–term groups. The central zone of the chart, with lines crossing each other, shows the turning point of this trend, where legal terms outnumber the environmental ones. Moreover, Figure 2 reveals a quite homogeneous distribution of terms which can refer to both domains (referred to as ‘Environmental/Legal’ in Table III). It is the case of terms such as *politica ambientale* ‘environmental policy’, *obiettivo ambientale* ‘environmental object’, *informazione ambientale* ‘environmental knowledge’, etc. which have been categorized by both domain experts as belonging to a ‘twilight’ zone since they express general legal concepts which acquire a domain–specific meaning. Interestingly, the analysis carried out by the legal expert highlighted that some of the acquired environmental terms are explicitly defined in the legal texts being considered: such terms are associated with a high contrastive score and are located in the first 30–term group. This is the case of *rifiuto pericoloso*, ‘hazardous waste’, *sostanza pericolosa*, ‘hazarous substance’, *valore limite di emissione*, ‘emission limit value’, etc. whose meanings are explicitly defined in the acquisition corpus. For example, Article 2 “Definitions”, letter g) of the *Regulation (EC) no 2150/2002 of the European Parliament and of the Council of 25 November 2002 on waste* *statistics* contains the following definition of 'hazardous waste': "hazardous waste shall mean any waste as defined in Article 1(4) of Council Directive 91/689/EEC of 12 December 1991 on hazardous waste". It may be possible to conclude that such terms are particularly relevant for the regulated domain being considered, and for this reason, occur with higher frequencies in the target domain. This could open interesting developments in the field of legal re–definition of the regulated–domain terms. In fact, as overtly pointed out in Walter et al. (2006), the successful retrieval of definitions contained in statutes and legal texts can help providing a large knowledge base to be used in text–based ontology learning tasks.

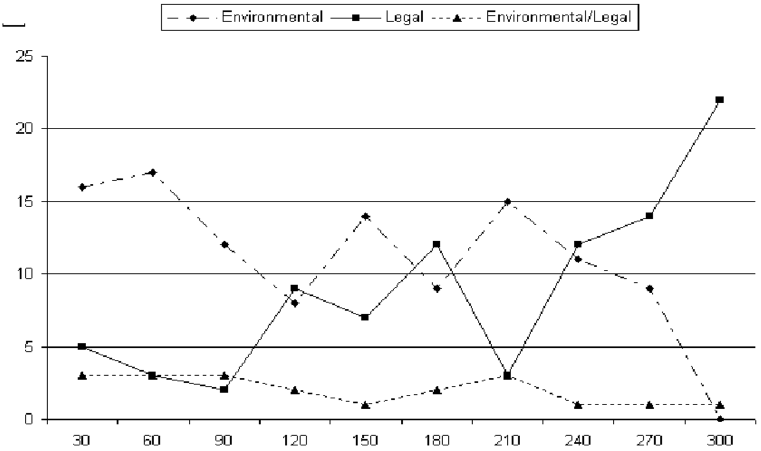

Figure 2. Distribution of the three types of terms in the extracted multi–word term list

6. Conclusion

In this paper, we showed how a modular and contrastive approach to term extraction can be usefully exploited in the legal domain to tackle the well–known *epistemological promiscuity* problem. To our knowledge, it is the first time that such a problem has been addressed in the terminology extraction literature with successful results. In the proposed modular approach to term extraction, candidate single and multi–word terms are first identified using state–of–the–art statistical measures and are subsequently filtered by applying a contrastive reranking method aimed at discriminating between acquired legal terms and regulated–domain terms. The evaluation of achieved results, carried out with the help of domain experts, showed that the proposed approach is really effective in dealing with particularly challenging text types, such as legislative texts.

7. Acknowledgments

The research reported in the paper has been partly supported by the Italian FIRB project "Piattaforma di servizi integrati per l'Accesso semantico e plurilingue ai contenuti culturali italiani nel web". The authors would like to thank Angela D'Angelo of the Scuola Superiore Sant'Anna of Pisa and Paolo Plini of the Institute of Atmospheric Pollution, Environmental Terminology Unit (CNR, Rome) who contributed as domain experts to the evaluation process.

References

Basili, R., Moschitti, A., Pazienza, M.T., and Zanzotto, F. (2001), *A contrastive approach to term extraction*, in Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA–2001), Nancy.

Biagioli, C. (2009), *Modelli funzionali delle leggi. Verso testi legislativi autoesplicativi*, Series in Legal Information and Communication technologies, vol. 6, European Press Academic Publishing.

Bonin, F., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2010), *A Contrastive Approach to Multi–word Term Extraction from Domain Corpora*, in Proceedings of the "7th International Conference on Language Resources and Evaluation (LREC 2010)", La Valletta, Malta, 19–21 May, pp. 3222–3229.

Breuker, J. and Hoekstra, R. (2004), *Epistemology and Ontology in Core Ontologies: FOLaw and LRI-Core, two core ontologies for law*, in Proceedings of the "Workshop on Core Ontologies in Ontology Engineering" (EKAW04), Northamptonshire, UK, pp. 15-27.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005) *Ontology Learning from Text: an Overview*, In Buitelaar et al. (eds.), *Ontology Learning from Text: Methods, Evaluation and Applications* Volume 123, Frontiers in Artificial Intelligence and Applications, pp. 3–12.

Casellas, N. (2008), *Modelling Legal Knowledge through Ontologies. OPJK: the Ontology of Professional Judicial Knoweldge*, Ph.D. thesis, Institute of Law and Technology, Autonomous University of Barcelona.

Chung, T.M., and Nation, P. (2004), *Identifying technical vocabulary*, in *System, 32*, pp. 251–263.

Dell'Orletta, F. (2009), *Ensemble system for Part-of-Speech tagging*, in Proceedings of "Evalita'09", Reggio Emilia, December.

Francesconi, E. (2010), *Legal Rules Learning based on a Semantic Model for Legislation*, in Proceedings of the "Workshop on Semantic Processing of Legal Texts" (SPLeT-2010), held in conjunction with the 7th Conference on Language Resources & Evaluation (LREC 2010) La Valletta, Malta, 23rd May, (in press).

Frantzi, K., and Ananiadou, S. (1999), *The C–value / NC Value domain independent method for multi–word term extraction*, in *Journal of Natural Language Processing*, 6(3), pp. 145–179.

Lame, G. (2005), *Using NLP techniques to identify legal ontology components: concepts and relations*, in Benjamins et al. (eds.), Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications, Lecture Notes in Computer Science, Volume 3369, pp. 169–184.