


Proceedings of LOAIT 2010

IV Workshop on Legal Ontologies and Artificial Intelligence Techniques

7 July 2010, Fiesole (Florence, Italy)

Enrico Francesconi
Simonetta Montemagni
Piercarlo Rossi
Daniela Tiscornia (Eds.)

Program Committee

- Gian Maria Ajani, University of Turin, Italy
- Tommaso Agnoloni, ITTIG-CNR, Italy
- Trevor J.M. Bench-Capon, University of Liverpool, UK
- V. Richard Benjamins, Telefónica R&D, Spain
- Guido Boella, University of Turin, Italy
- Alexander Boer, Leibniz Center for Law, University of Amsterdam, The Netherlands
- Joost Breuker, Leibniz Center for Law, University of Amsterdam, The Netherlands
- Thomas Bruce, Cornell Law School, US
- Paul Buitelaar, DERI research institute in Galway, Ireland
- Pompeu Casanovas, Institute of Law and Technology, Universitat Autònoma de Barcelona, Spain
- Nuria Casellas, Institute of Law and Technology, Universitat Autònoma de Barcelona, Spain
- Aldo Gangemi, Institute of Cognitive Sciences and Technologies (ISTC-CNR), Italy
- Roberto García, Universitat de Lleida, Spain
- Guido Governatori, NICTA, Queensland Research Laboratory, Australia
- Rinke Hoekstra, Leibniz Center for Law, University of Amsterdam, The Netherlands
- Mustafa Jarrar, Birzeit University, Palestine
- Michael Klein, VU University Amsterdam, The Netherlands
- Alessandro Lenci, Department of Linguistics, University of Pisa, Italy
- Monica Palmirani, University of Bologna, Italy
- Wim Peters, Natural Language Processing Research Group, University of Sheffield, UK
- Giovanni Sartor, European University Institute, Florence, Italy
- Marco Schorlemmer, IIIA-CSIC, Spain
- Erich Schweighofer, University of Vienna, Austria
- Barry Smith, University at Buffalo, US
- Pierluigi Spinosa, ITTIG-CNR, Italy
- York Sure, SAP Research, Germany
- Tom van Engers, Leibniz Center for Law, University of Amsterdam, The Netherlands
- Réka Vas, Department of Information Systems, University Corvinus of Budapest, Hungary
- Radboud Winkels, Leibniz Center for Law, University of Amsterdam, The Netherlands
- Adam Wyner, Department of Computer Science, University College London, UK

Contents

SECTION I

Legal Knowledge Extraction

Towards Annotating and Extracting Textual Legal Case Elements

Adam Wyner
University of Leeds

Abstract.
In common law contexts, legal cases are decided with respect to precedents rather than legislation as in civil law contexts. Legal professionals must find, analyse, and reason with and about cases drawn from a set of cases (a case base). A range of particular textual elements of a case may be relevant to query and extract. Commercial providers of legal information allow legal professionals to search a case base by keywords and meta data. However, the case base and the search tools are proprietary, of limited, non-extensible functionality, and are restricted access. Moreover, no provider applies natural language processing techniques to the cases for text analysis, XML annotation, or information acquisition. In this paper, we discuss an initial experiment in developing and applying natural language processing tools to cases to produce annotated text which can then support information extraction.

Keywords: Text Analysis, Legal Cases, Ontologies

1. Introduction

In common law contexts, judges and juries decide a legal case to follow previously decided cases (precedents) rather than legislation as in civil law contexts.[1] The set of such cases is the legal case base. Legal professionals must find, analyse, and reason with and about cases drawn from the case base in the course of arguing for a decision in a current undecided case. A range of elements of cases may be relevant to query and extract such as the citation index, participants, locale, jurisdiction, representatives, judge, prototypical fact patterns (factors), applicable law, and others. Commercial providers of legal information allow legal professionals to search the case base by keywords and meta data. However, the case base and search tools are proprietary, of limited, non-extensible functionality, and are restricted access. Moreover, no provider works with Semantic Web functionalities such as ontologies or rich XML annotations, nor are natural language processing techniques applied to the cases to support analysis to acquire information.

Text annotation of unstructured linguistic information is a significant, difficult aspect of the "knowledge bottleneck" in legal information processing. In this paper, we apply natural language processing tools to textual elements in cases, which are unstructured text, to produce annotated text, from which information can be extracted, thus contributing to overcoming the bottleneck. The extracted information can then be submitted to further processes.

[1] Correspondence to Adam Wyner adam@wyner.info.

Where the annotations are associated with an ontology (Wyner and Hoekstra, 2010) along with an associated case based reasoner (Wyner and Bench-Capon, 2007), then we make progress towards a textual case based reasoning system which enables processing from natural language case decisions in the case base to generated decisions in novel cases (Weber et al, 2005a). However, this paper focuses on the initial development in annotating cases with respect to case elements.

The paper is a feasibility study for future research on information extraction of case elements. [2] In this paper, we focus on case elements rather than case factors (see (Wyner and Peters, 2010)).

In 2, we discuss background and materials. In 3, we present the methodology, which uses the General Architecture for Text Engineering(GATE) system, sample components of system, sample results, and a work flow for further refinement.[3] Finally, in 4, we review the paper and outline future work to evaluate and improve our results.

2. Background and materials

Legal case based reasoning with factors has been a topic of central concern in artificial intelligence and law. For our purposes, there are two main branches of research. One branch, knowledge representation and reasoning systems, requires a knowledge base that is constructed by manual analysis (cf. (Hafner, 1987), (Ashely, 1990), (Rissland et al, 1996), (Aleven, 1997), (Wyner and Bench-Capon, 2007)). However, this branch of research does not address the knowledge bottleneck, which is the extraction of information to compose the knowledge base.

The other branch, information extraction, addresses the bottleneck using natural language processing techniques which identify informative components of the text and annotate them with XML. The annotated information can be extracted with *XQuery*. Thus, the content of the documents can be identified from its source linguistic realisation. There are a range of areas where information extraction of legal texts has been carried out: ontology construction ((Lame, 2004) and (Peters, 2009)), text summarisation ((Moens et al, 1997) and (Hachey and Grover, 2006)), extraction of precedent links (Jackson et al, 2003), and factor analysis ((Ashley and Brüninghaus, 2009) and (Wyner and Peters, 2010)). We focus on information extraction of case elements, which contributes to this previous work.

The branches are related since the extracted information can be represented in some knowledge base and reasoned with. For case based reasoning with factors as in (Aleven, 1997), we extract factors; for reasoning about

[2] Contact the author for materials.

[3] For GATE, see http://gate.ac.uk/.

precedential relations among cases (overturned, affirmed, and so on), we extract citation indices and relational terms. As legal cases are not just about the law *per se*, but about some content area (e.g. intellectual property, family law, etc) and human properties and artifacts (e.g. instruments and property), one might suppose that all of human knowledge and experience is potentially under the scope of the law and so potentially to be extracted, put in a knowledge base, and reasoned with (cf. works on legal knowledge representation (Peters et al, 2007), (Scheighofer and Liebwald, 2007), (Hoekstra et al, 2009), and (Gangemi et al, 2005)). Yet, (Wyner and Hoekstra, 2010) argue that the focus should be on information which has a legal definition or function, leaving aside high level, non-legal domain information (e.g. events/processes, causation, time, and so on).

In this light and in the current paper, we are interested in case information that would be relevant to searching for or extracting information from cases. For reasons of space, we only give a sample of the information we searched for and annotated:

- Case citation, cases cited, precedential relationships.
- Names of parties, judges, attorneys, court sort....
- Roles of parties, meaning plaintiff or defendant, and attorneys, meaning the side they represent.
- Final decision.

With respect to these features, one would want to make a range of queries (using some appropriate query language) such as:

- In what cases has company X been a defendant?
- In what cases has attorney Y worked for company X, where X was a defendant?

As we initially based our work on information extraction from California Criminal Courts in (Bransford-Koons, 2005), developing and modifying lists and rules, we worked with a legal case base of cases from the United States. (Bransford-Koons, 2005) reports working with 47 criminal cases drawn from the California Supreme Court and State Court of Appeals. However, only two cases are given as samples and for which we have access; for this feasibility study, we give examples from these cases. (Bransford-Koons, 2005) uses GATE (described below) and OPENCYC, which is a repository of common sense rules. We do not consider OPENCYC here. To show the feasibility of the approach, we provide preliminary results on this very small corpus of *People v. Coleman 117 Cal.App.2d 565* and *In re James M., 9 Cal.3d 517*.

3. Methodology using GATE

We use the GATE framework (Cunningham et al, 2002). GATE Developer is an open source desktop application written in JAVA and for linguists and text engineers. Using a GUI, it allows a variety of text analysis tools to be cascaded and applied to a set of documents.

For our purposes, we have applied natural language processing modules such as Tokeniser, Gazetteer, and Java Annotation Patterns Engine (JAPE), each module providing input to the next. The last two modules are explained further below.

In addition to these functionalities, one can also use entity extraction and syntactic parsing components. For a particular domain, it is important to provide gazetteer lists and JAPE rules. In general, there is a cascade from *lower level* information in the parts of speech and gazetteer lists to *higher level* information where lower level information is used to compose more complex units of information. As a working strategy, the lists capture simple, unsystematic patterns, leaving the JAPE rules to capture systematic, complex patterns.

Figure 1 represents the work flow (derived from the work flow diagram in (Wyner and Peters, 2010)), where an initial specification guides the definition of gazetteer lists and JAPE rules. The process cascade is applied to the corpus, which results in an annotated text. Examining the results, one determines what to modify in the gazetteer lists and JAPE rules until one achieves desired annotations. Thus, we have an *iterative process* which supports experimental refinement of the lists and rules that induce annotation.

3.1. GAZETTEER LISTS

A gazetteer is a list of lists. Each list is comprised of strings that are associated with a central concept or with some elements of the text. The lists annotate the words and strings with the MajorType of the list; they provide the bottom level of annotation on which higher level annotations are constructed using JAPE rules. The gazetteer lists discussed here are manually composed.

We initially worked with gazetteer lists from (Bransford-Koons, 2005). However, while the lists may "work", they are clearly in need of reconstruction and extension, which we discuss. One observation is that the lists are defined for US case law and particularly the California district courts. Thus, we cannot simply apply the lists to different jurisdictions, e.g. the United Kingdom; the lists and rules must be localised to different contexts. For instance, the term `Fifth Appellate District` or `Municipal Court of....` may not occur in the UK. Similar issues arise with case citations, roles of participants, causes of action, and so on. More technically, lists have alternative graphical (capital or lower case) or morphological forms, which would be bet-

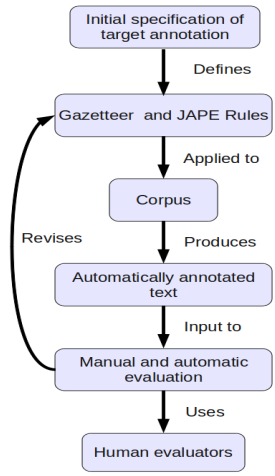

Figure 1. A Workflow Diagram

ter addressed using GATE's Flexible Gazetteer, which homogenises graphical forms and lemmatises words (providing a "root" form). As a general strategy, it is best to create lists with "unique" word forms or fixed phrases rather than those which may otherwise be constructed by JAPE rules. Taking these considerations into account, we created lists for particularly legal terminology and used the Flexible Gazetteer. The lists thus comprise a conceptual cover term; for example, a search for judgments or legal parties in a corpus will return cases and passages which contain terms found in these lists:

- judgements.lst. Terms related to judgment: grant, deny, reverse, overturn, remand,....
- legal_parties.lst. Terms for legal roles: amicus curie, appellant, appellee, counsel, defendant, plaintiff, victim, witness,....

A range of lists such as the two sampled below bear on "indicators" of structure. For example, "v." is used in cases to indicate the opposing parties, so it can be used to leverage identification and annotation of parties which appear on either side of the indicator. These are not unproblematic: the indicator might incorrectly label an abbreviated first name. There may be better ways to find judges than the initial "J."; in particular, as the list of judges is finite and give by the court system, it might be simplest to use such a list rather than applying text mining to finding it.

- legal_casenames.lst. Terms that can be used to indicate case names: v., In Re,....
- judgeindicator.lst. The indicator J.. This is a problematic indicator if it is part of an individual's name.

In other lists, we have phrases, abbreviations, and case citations. For phrases, there are two strategies. (Bransford-Koons, 2005) follows the strategy of listing the possible phrases. The alternative which we adopt is to provide bottom level lists for constituent parts of the phrases, then constructing the complex phrases by rule. The former requires a finite list; it will not annotate a novel phrase. Constructing phrases requires that the output be checked against actual phrases so it does not over generate. The treatment of abbreviations in GATE is not entirely clear, though (Bransford-Koons, 2005) simply lists them. For example, one would want to link the abbreviation with the full form, e.g. `Fifth Appellate District` and `Fifth App. Dist.`, and moreover, there may be a range of alternative abbreviations. One strategy is to have related lists - a list of phrases where the abbreviation of the phrase is a `MinorType`, and a list of abbreviations where the correlated phrase is a `MinorType`. In our view, more general solutions are better than specific ones which list information; lists ought to be contain arbitrary information, while JAPE rules construct systematic information. Case citations combine the issues of phrases, abbreviations, and alternative forms. We may have a citation such as `Cal.App. 3d` which abbreviates the California Court of Appeals, Third District. Clearly, each part is a component that can be reused in other citations. Moreover, as spaces matter in text analysis, we must account for alternatives, `Cal.App.3d` and `Cal. App. 3d`.

- lower_courts.lst. Phrases for other courts: Municipal Court of, Superior Court of,....
- legal_code_citations.lst. Code citations: Civ. Code, Penal Code,....

Some of the terms are functional; that is, both legal parties and counsel names are roles that individuals have with respect to a particular context. In one context, an individual may be a plaintiff, while in another the defendant. In annotating an individual with a functional role, e.g. an individual as plaintiff, we rely on local context within the text and do not presume that the individual's annotation applies across cases.

Finally, (Bransford-Koons, 2005) provides a range of terms which relate to the content of the case. For example, a case of criminal assault is marked by the appearance of terms bearing on weapon or intention.

- weapons.lst. A list of items that are weapons: assault rifle, axe, club, fist, gun,....

- intention.lst. Terms for intention: intend, expect,....

While it would be meaningful to index cases according to such content, they present several problems. Clearly, whether something is a weapon or criminal assault is context dependent since in some other context they might not be. How could one bound the range of relevant terms appropriately and give them interpretations that are relevant to the context? For example, isn't any object a possible weapon? These may be terms which, as discussed in (Wyner and Hoekstra, 2010), are developed in independent modules; we do not want to develop a full theory of space, time, instruments, intention, or causation.

3.2. JAPE RULES

Given the bottom-level annotations provided by the lists, we have JAPE rules which make the annotations graphically represented and available for higher level annotations. Below is a partial list of annotations given by JAPE rules.

- AppellantCounsel: annotates the appellant counsel.
- DSACaseName: annotates the case name.
- CauseOfAction: annotates for causes of action.
- DecisionStatement: annotates a sentence as the decision statement.
- JudgeName: annotates the names of judges.

Some of the JAPE rules simply translate the Lookup type into an annotation such as `Weapon`, while other rules use the Lookup type and context to annotate a text span such as `AppellantCounsel` and `DecisionStatement`. In the following sample rule, a sentence which contains a judgment term (e.g. affirm, overturn, etc) followed by a judge's name is labeled a decision statement. The rule relies on a standard format, where the case decision is followed by the judge's name; were similar patterns to appear in the case, then they too might be mis-annotated as a decision of the case.

```
Rule: DecisionStatement
Priority: 10
(
{Sentence contains JudgementTerm}
):termtemp
{JudgeName}
–>
:termtemp.DecisionStatement = {rule = "DecisionStatement"}
```

3.3. RESULTS

In this section, we give some of the results of running our GATE application over our corpus, giving the results using the graphical output of GATE

We have the following sample outputs from our lists and rules applied to *People v. Coleman, 117 Cal App. 2d 565*. The coloured highlights on the case text are associated with the same coloured annotation. We can output an XML representation to indicate the annotation. In Figure 2, we find the address, court district, citation, case name, counsels for each side, and the roles. The results give a flavour of the annotations, though further work is required to refine them.

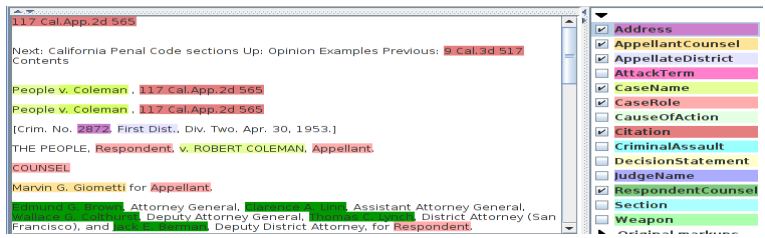

Figure 2. Case Information I

In Figure 3, we focus on additional information such as structural sections (e.g. Opinion), the name of the judge, and terms having a bearing on criminal assault and weapons. In Figure 4, we identify the decision.

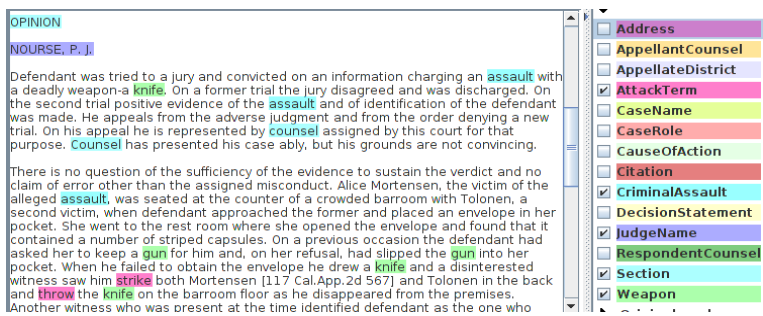

Figure 3. Case Information II

Figure 4. Case Information III

4. Conclusion

In this paper, we have outlined and extended a proof of concept approach to text mining legal cases in order to extract a range of particular elements of information from the cases. While a relatively small system applied to a very small corpus, the lists and rules approach can be extended further and relatively easily. Further developments using this approach to text mining would be to relate the extracted information to an ontology which is directly incorporated into the GATE pipeline. A second development would be to engage a wide range of users (e.g. law school students) in a collaborative, on line annotation task using GATE TeamWare. Not only would this have didactic purposes (to focus the attention of students on close analysis of the text), but it would also help to build up a body of annotated texts for further research as well as development of a gold standard that could be used for machine learning.

References

Aleven, A. (1997), *Teaching case-based argumentation through a model and examples*. PhD thesis, University of Pittsburgh, 1997.

Ashley, K. (1990), *Modelling Legal Argument: Reasoning with Cases and Hypotheticals*. Bradford Books/MIT Press, Cambridge, MA, 1990.

Ashley, K. and Brüninghaus, S. (2009), Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law*, 17(2):125–165, 2009.

Bransford-Koons, G. (2005), Dynamic semantic annotation of California case law. Master's thesis, San Diego State University, 2005.

Cunningham, H., Maynard, D., Bontcheva, K, and Tablan, V. (2002), GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

Gangemi, A., Sagri, M., and Tiscornia, D. (2005), A constructive framework for legal ontologies. In V.R. Benjamins, P. Casanovas, J. Breuker, and A. Gangemi, editors, *Law and the Semantic Web*, pages 97–124. Springer Verlag, 2005.

Hachey, B. and Grover, C. (2006), Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345, 2006.

Hafner, C. (1987), Conceptual organization of case law knowledge bases. In *ICAIL '87: Proceedings of the 1st International Conference on Artificial Intelligence and Law*, pages 35–42, New York, NY, USA, 1987. ACM.

Hoekstra, R., Breuker, J., Bello, M., and Boer A. (2009), LKIF core: Principled ontology development for the legal domain. In Joost Breuker, Pompeu Casanovas, Michel C. A. Klein, and Enrico Francesconi, editors, *Law, Ontologies and the Semantic Web*, volume 188 of *Frontiers in Artificial Intelligence and Applications*, pages 21–52. IOS Press, 2009.

Jackson, P., Al-Kofahi, K., Tyrell, A., and Vachher, A. (2003), Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150(1-2):239–290, November 2003.

Lame, G. (2004), Using NLP techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, 12(4):379–396, 2004.

Moens, M.-F., Uyttendaele, C., and Dumortier, J. (1997), Abstracting of legal cases: the salomon experience. In *ICAIL '97: Proceedings of the 6th International Conference on Artificial Intelligence and Law*, pages 114–122, New York, NY, USA, 1997. ACM.

Peters, W. (2009), Text-based legal ontology enrichment. In *Proceedings of the workshop on Legal Ontologies and AI Techniques*, Barcelona, Spain, 2009.

Peters, W., Sagri, M.-T., and Tiscornia, D. (2007), The structuring of legal knowledge in LOIS. *Artificial Intelligence and Law*, 15(2):117–135, 2007.

Rissland, E., Skalak, D., and Friedman, T. (1996), BankXX: Supporting legal arguments through heuristic retrieval. *Artificial Intelligence and Law*, 4(1):1–71, 1996.

Schweighofer, E. and Liebwald, D. (2007), Advanced lexical ontologies and hybrid knowledge based systems: First steps to a dynamic legal electronic commentary. *Artificial Intelligent and Law*, 15(2):103–115, 2007.

Weber, R., Ashley, K., and Brüninghaus, S. (2005), Textual case-based reasoning. *Knowledge Engineering Review*, 20(3):255–260, 2005.

Wyner, A. and Bench-Capon, T. (2007), Argument schemes for legal case-based reasoning. In Arno R. Lodder and Laurens Mommers, editors, *Legal Knowledge and Information Systems. JURIX 2007*, pages 139–149, Amsterdam, 2007. IOS Press.

Wyner, A. and Hoekstra, R. (2010), A legal case OWL ontology with an instantiation of *Popov v. Hayashi*. *Knowledge Engineering Review*, xx:xx, 2010. To appear.

Wyner, A. and Peters, W. (2010), Towards annotating and extracting textual legal case factors. In *Proceedings of the Language Resources and Evaluation Conference Workshop on Semantic Processing of Legal Texts*, Malta, 2010. To appear.

Suggesting Model Fragments for Sentences in Dutch Law

Emile de Maat*, Radboud Winkels*
*Leibniz Center for Law, University of Amsterdam, {demaat|winkels}@uva.nl

Abstract. A main issue in the field of artificial intelligence and law is the translation of source of law that are written in natural language into formal models of law. This article describes a step in that transformation: the creation of models for individual sentences in a source of law. The approach uses a natural language parse to analyse the sentence, and then translates the resulting parse tree to a formal model, using both generic and law-specific attributes.

Keywords: Automated Modelling, Natural Language Processing

1. Introduction

A main issue in the field of artificial intelligence and law is the transformation of sources of law that are written in natural language (and therefore rather informal) into formal models of law that computers can reason with. This is a time and effort consuming process, error prone and different knowledge engineers will arrive at different models for the same sources of law. Moreover, these models should be closely linked to the original sources (and at the right level of detail, i.e. isomorphic) since these sources tend to change over time and maintenance of the models is a serious problem. This calls for tools and a method for supporting this modelling process and increasing inter-coder reliability.
We have been researching a method to create isomorphic models semi-automatically, focusing on (Dutch) laws. This article presents a next step in this creation process.

1.1. General Approach

In order to achieve (semi-)automatic modelling of legal sources, we follow a number of steps, as shown in figure 1. The process starts with the source document, written in natural language (Dutch). Currently, we focus on laws, though we hope to expand to other types of legal sources later on. We first make the structure of the document explicit, by marking up the different parts, such as chapters, paragraphs and sentences, and assigning identifiers to each part. We then proceed to mark all references to other legal sources that are contained in the text, using a parser based on patterns for references (see (de Maat, 2006)). This structure and reference information is stored in CEN/MetaLex XML[1].

[1] See http://www.metalex.eu/

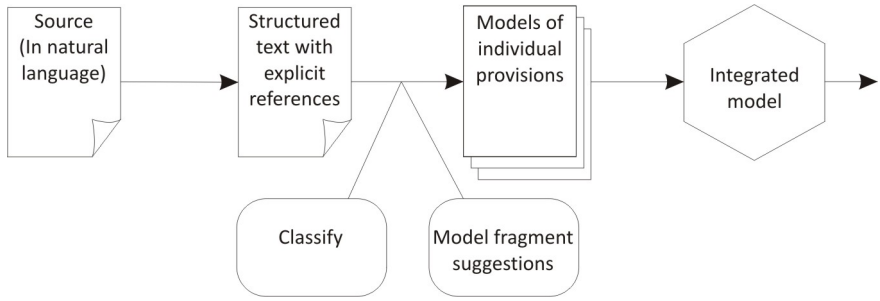

Figure 1. Steps in automatic modelling of legal texts.

The next step is to create models for each individual statement in the text. In most cases, each sentence in Dutch law forms a complete statement (though possibly part of a bigger construct), so we are, in fact, creating a model for each sentence in the text. In the last step, these individual models are integrated with each other to come to a complete model. In order to create the models, we start by classifying each sentence in the text as a specific provision, such as a definition, a duty, or a modification of an earlier law. In total, we recognise ten different main categories. As with the references, this is done by recognising certain patterns in the text (de Maat, 2008).
For several types of statements, such as modifications and setting the enactment date or citation title, recognising the pattern and classifying the sentence is also nearly sufficient for creating a model of the sentence. For example:

Aliens Act 2000
This law is referred to as: Aliens Act 2000.

This sentence is classified by the pattern "is referred to as", which splits the sentence in two parts: a reference (recognised by the reference parse) to "this law" and a citation title. This is all the information that is needed to represent the meaning of this sentence[2]. More elaborate sentences, that contain terms relating to the subject matter that the law is about, require more detailed analysis[3]. A natural language parser can provide such a more detailed analysis. This paper describes our initial experiences while using a natural language parser to enhance the input for our modeller. For this research, we have used the Alpino parser for Dutch (Bouma, 2001) to parse the sentences. The Alpino parser assigns a dependency structure to the sentence. These structures are described by Bouma et al:

[2] As said, this also holds true for sentences containing modifications to other legal sources. However, for such sentences, analysis of the modified text is needed to determine the full impact (not meaning) of such a sentence.

[3] This applies to norms, definitions and many application provisions. Earlier research (de Maat, 2008) suggests that these comprise about 64% of the sentences encountered.

Dependency structures make explicit the dependency relations between constituents in a sentence. Each non-terminal node in a dependency structure consists of a head-daughter and a list of non-head daughters, whose dependency relation to the head is marked.

The dependency structure can be stored as an XML file, which is the format we use as input for our modeller.

2. Creating Model Fragments

Our approach is similar to that published in (Biagioli, 2005), where Italian laws are modelled. However, Biagioli et al. aim for fairly rough frames; for example, for an obligation, their approach attempts to fill the slots addressee, action and third-party. We hope to achieve some more detail, splitting up these fields in more parts. In this sense, our method comes closer to those of (Sarwar Bajwa, 2009), who generates UML models from parse trees, (McCarty, 2007), who transforms parse trees to quasi-logical form, or (Bos, 2004), who translate parse trees to first order logic statements. Both these methods map individual words to model elements. An example by Bos et al:

The school-board hearing at which she was dismissed was crowded with students and teachers.

This results in the following first-order logic statement:

$$\exists a((school-board(a) \wedge hearing(a)) \wedge \exists b(female(b) \wedge \exists c(dismiss(c) \wedge (patient(c,b) \wedge (at(a,c) \wedge \exists d(crowd(d) \wedge (patient(d,a) \wedge (\exists e(student(e) \wedge with(d,e)) \wedge \exists f(teacher(f) \wedge with(d,f))) \wedge event(d)))))))))$$

We wish to mix these approaches. For normative sentences, this means that we see each normative sentence as describing a situation that is allowed or disallowed. We consider the main verb of a sentence as the action that is allowed or disallowed, with the other elements being modifiers or properties of that action. For example:

Our Minister issues a warrant to the negligent person.

The main verb of this sentence is *to issue*, so that is considered the action. Properties of this action are the subject (*Our Minister*), the direct object (*a warrant*) and the indirect object (*the negligent person*). All these elements are distinguished by the Alpino parser, allowing us to extract them for our model. Within Dutch law, this sentence format expresses an obligation, so the action as a whole is classified as an obligation.

Obligation	
Action	issue
Subject	Our Minister
Direct Object	warrant
Indirect Object	negligent person

The articles (*the*, *a*) are left out of the model, though they are stored internally, as they are of importance during later integration of the model; *the negligent person* often is a reference to an earlier sentence, whereas *a negligent person* is not.

Further detail can be added by splitting of adjectives and relative clauses from the noun they modify. For example, *negligent person* has two properties: being a person and being negligent. Splitting adjectives from nouns is not always desirable; it is preferable to leave multiword expressions intact. *European Union* is not any union that is also European; *Our Minister of Finance* is not any minister that is also ours, and of finance[4]. Instead, these are references to concepts that have been defined elsewhere: the common sense domain, the juridical domain or elsewhere in this law. Common multiword expressions are recognised by the Alpino parser; juridical domain or law-dependent expressions need be filtered out separately.

Relative clauses are more complex then adjectives, as they contain a complete new sentence. In this case, we repeat the procedure for the main sentence, identifying the main action and all properties of that action. For example:

Our Minister issues a warrant to the person that neglected his duties.

This sentence would yield a frame like[5]:

Obligation		
Action	issue	
Subject	Our Minister	
Direct Object	warrant	
Indirect Object	person	
	subjectOf	
	Action	neglect
	Direct Object	his duties

[4] In Dutch laws, *Our Minister of Finance* is a reference to the (Dutch) Minister of Finance. No more detailed model is needed, as no derivations need to be made.

[5] For the moment we use a frame-like representation. These look somewhat like the frames presented by (van Kralingen, 1995), but these were more legally oriented and had a fixed number of slots, while our structures are more dynamic and language oriented

2.1. Filtering out signal words

The sentences we showed above are examples of normative sentences that do not use signal words; only the desired situation is described, and it is left implicit that this is an obligation. Other sentences in the law use signal words to make the kind of norm explicit, such as:

The buyer is obliged to pay the price.[6]

This sentence uses *is obliged* to make it clear that this is an obligation. Other examples of signal words are *must*, *may* and *is allowed*. These sentences require a different approach than the sentences without signal words. If we were to use the same approach, the result would be something like:

Obligation	
Action	is obliged to pay
Subject	buyer
Direct Object	price

This is not a desirable outcome, as the action that this norm deals with is pay rather than is obliged to pay. When modelling these sentences, these signal words should not be included in the model of the situation (their meaning is translated into whether the situation is allowed or disallowed). Ideally, after we have categorised the sentence (based on the signal words), we would like to transform the sentence to a sentence without signal words, like:

The buyer pays the price.

We could then model that sentence to come to a correct frame. Simply leaving out the signal words may lead to errors, since the role of the other words might need to shift as well. However, the parse of the sentence actually contains this "transformed sentence" that we want to model. This is shown in figure 2.

Beneath the body node, we find exactly the sentence that we are looking for. Alpino assigns this dependency structure to any sentence that follows this pattern. This makes it easy to filter out the signal words by simply focusing on the part of the parse tree that contains the transformed sentence. For each pattern we use for classification, it seems possible to define a part of the parse tree that should be ignored in order to come up with a correct model.

2.2. Passive voice

Many sentences in Dutch law are phrased in the passive voice, such as this instruction:

[6] Dutch Civil Code, BW7, article 6 sub 1.

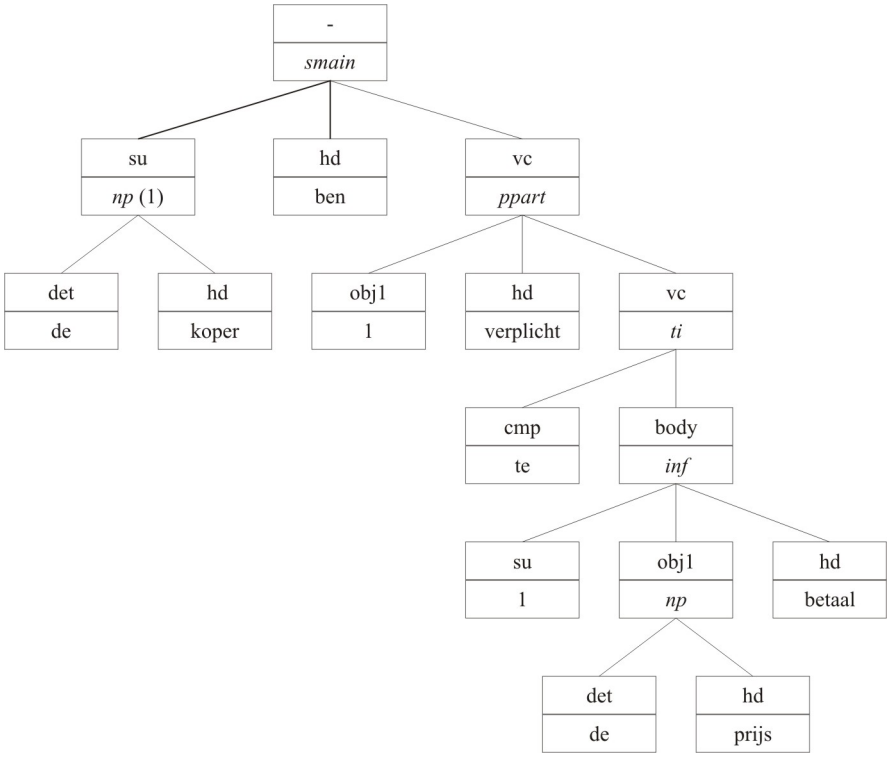

Figure 2. Alpino parse tree (with reduced information) for "The buyer is obliged to pay the price" (in Dutch).

An English translation is added to this report.[7]

A sentence in the passive voice cannot be modelled in the same way as a regular sentence, as the subject of the sentence is actually the direct object, and should be modelled as such. Again, the parse of the sentence gives us an easy way to do this:

The verb clause (vc) of the sentence holds the sentence in active voice, with the subject re-cast in the role of object. By modelling the verb clause instead of the sentence as a whole, we get the correct model, with the correct object, and without the auxiliary verb. If the actual subject is present in the sentence (for example, if the sentence would read*An English translation is added to this report by the organiser*), then this prepositional object is not re-cast in the role of object in the tree. We will have to detect its presence by scanning for signal words like *by*. As this does not always indicate a subject, this will be one of the cases were human validation is necessary.

[7] Law for the protection of Antarctica, article 33, sub 3

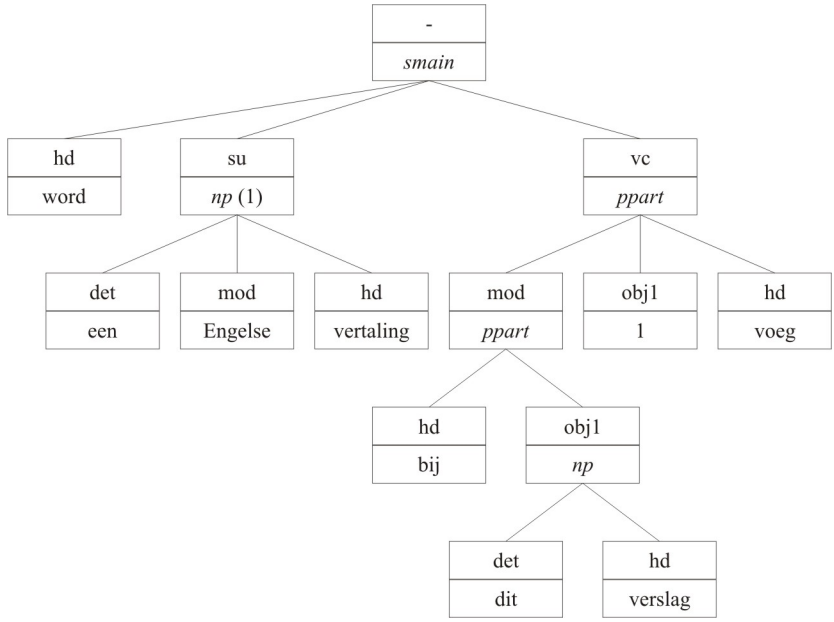

Figure 3. Alpino parse tree (with reduced information) for 'An English translation is added to this report" (in Dutch).

2.3. LISTS

Lists are also recognised by the Alpino parser, and can therefore easily be added to our models as the union or intersection of the different list items, depending on the conjunction used. However, though the conjunction *and* suggests an intersection, it often expresses a union instead. For example:

Advances and duties are paid in cash.

In this sentence, it is the union of *advances*and *duties*that is meant. Our current approach is to translate *and*with a union if it appears in a relative clause, and with an intersection otherwise.

2.4. NEGATION

Negative sentences should also be recognised, and modelled as the "positive" sentence, with the additional notion that it is inverted. This can usually be done by not including certain signal words as element in the model, but by inverting the model if it is encountered. The most common signal word is *not*. If it is encountered, it is not added to the frame, but instead, the containing element is marked as inverted. The determiner*no* is another example of a signal word for negation. However, it can affect more than its containing element. For example:

No bodies are interred on a closed cemetery.

This is an obligation, and the direct object of this sentence is *no bodies*. However, if we apply the negation simply to the object, i.e. the object is "not a body", it would imply the obligation of to bury something that is not a body on the cemetery. Instead, we need to apply the negation to the entire sentence: On is obliged not to bury bodies at a closed cemetery.

3. Experiences

At this moment, we do not have a fully automated process to create the models, and have not yet tested this method on a large body of sentences. Instead, random sentences have been selected, parsed using Alpino and then fed into our modeller.

There is a clear difference between the computer generated models and those created by a human expert with regard to the granularity of the model. Our method will create models with model elements that represent one word from the original sentence, whereas a human expert is more likely to include some sentence fragments as a whole. For example, one Dutch law defines an alcoholic drink as *the drink that, at a temperature of twenty degrees Celsius, consists of alcohol for fifteen or more volume percents, with the exception of wine*. Our algorithm will dissect this sentence, whereas most human modellers will leave the first subordinate sentence intact and add it to the model as a single attribute (most likely abbreviated to *alcohol by volume*). A more detailed model seems not necessarily wrong, but quite possibly over-the-top and inconvenient for many applications.

The method assigns rather broad categorisations to each object (it is either a direct, indirect or prepositional object), but does not yet assign a legal meaning to such an object. It may be a third party involved or the instrument. Perhaps this is not an obstacle; users dealing with a system based on such models are likely to recognise the roles from the context and language used, whereas a computer does not need this information for the derivations we currently want to make. For future projects, though, the information may be required, and some way to automatically recognise it is desired.

For the modelling of norms, we have been focussing on the sentences that represent an obligation, duty or right. For those sentences, the method seems adequate. However, for other types of sentences, such as delegation, we have not come to an acceptable approach yet. Dealing with these sentences will require first of all that we recognise them. Currently, our classifier distinguishes only between obligation/prohibition and right/permission. Several of the patterns used clearly indicate delegations, but we have not yet established whether these patterns cover all delegations in Dutch laws.

A minor problem with regard to the parses made by Alpino is that most often,

the correct parse is not the one preferred by Alpino, but second, third or fourth. If we make several suggestions (each suggestion based on a parse by Alpino), this means that it will often not be the first suggestion that is correct, which means more effort is needed by a human expert who is verifying the models.
We expect that by expanding the lexicon used by Alpino, and perhaps by recalibrating the disambiguation on a written legal corpus, these problems will disappear.

4. Conclusion

We have presented a next step towards a method and tools for supporting the semi-automatic modelling of sources of law, necessary for an efficient, effective, and more reliable and pragmatic use of knowledge technology in the legal domain. We were already able to reliably detect structure in sources of law, find and resolve references in and between them, and classify individual sentences. Now we are able to suggest formal model fragments for certain types of the classifications. Though we are convinced that these model fragments will be a useful in supporting human experts creating models, we do feel that the approach is still too general. A more elaborate method is needed to create appropriate model fragments for different subtypes of sentences. Some method to avoid too granular models is desirable as well.

References

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S. and Soria, C. (2005), *Automatic semantics extraction in law documents*, in "Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL '05)", ACM, New York, pp. 133-140.

Bos, J., Clark, C., Steedman, M., Curran, J.R. and Hockenmaier, J. (2004), *Wide-Coverage Semantic Representations from a CCG Parser*, in 'Proceedings of the 20th international conference on Computational Linguistics", pp. 1240-1246.

Bouma, G., Noord, G. van, and Malouf, R. (2001), *Alpino: Wide Coverage Computational Analysis of Dutch*, in Daelemans, W., Sima'an, K., Veenstra, J. and Zavrel, J. (Eds.), "Computational Linguistics in the Netherlands CLIN 2000. Selected Papers from the Eleventh CLIN Meeting.", Rodopi, Amsterdam, pp. 45-59.

Kralingen, R. W. van (1995), *Frame-based Conceptual Models of Statute Law*, PhD thesis, Kluwer Law International, The Hague.

Maat, E. de, Winkels, R. and Engers, T. van (2006), *Automated Detection of Reference Structures in Law*, in Engers, T.M. van (Ed.), "Legal Knowledge and Information Systems. Jurix 2006: The Nineteenth Annual Conference", IOS Press, Amsterdam, pp. 41-50.

Maat, E. de and Winkels, R. (2008), *Automatic Classification of Sentences in Dutch Laws*, in Francesconi, E., Sartor, G. and Tiscornia, D. (Eds.), "Legal Knowledge and Information Systems. Jurix 2008: The Twenty-First Annual Conference", IOS Press, Amsterdam, pp. 207-216.

McCarty, L.T. (2007), *Deep semantic interpretations of legal texts*, in "Proceedings of the 11th International Conference on Artificial Intelligence and Law", ACM Press, New York, pp. 217-224.

Sarwar Bajwa, I., Samad, A. and Mumtaz, S. (2009), *Object Oriented Software Modeling Using NLP Based Knowledge Extraction*, European Journal of Scientific Research, Vol. 35, No. 1, pp. 22-33.

Multilingual Text Classification through Combination of Monolingual Classifiers

Teresa Gonalves, Paulo Quaresma
Departamento de Informtica, Universidade de vora
7000-671 vora, Portugal
*(*tcg@di.uevora.pt, pq@di.uevora.pt*)*

Abstract. With the globalization trend there is a big amount of documents written in different languages. If these polylingual documents are already organized into existing categories one can deliver a learning model to classify newly arrived polylingual documents. Despite being able to adopt a simple approach by considering the problem as multiple independent monolingual text classification problems, this approach fails to use the opportunity offered by polylingual training documents to improve the effectiveness of the classifier. This paper proposes a method to combine different monolingual classifiers in order to get a new classifier as good as the best monolingual one having also the ability to deliver the best performance measures possible (precision, recall and F_1). The proposed methodology was applied to a corpus of legal documents – from the EUR-Lex site – and was evaluated. The obtained results were quite good, indicating that combining different mono-lingual classifiers may be a promising approach to reach the best performance for each category independently of the language.

Keywords: Multilingual text classification, Machine Learning, Support Vector Machines

1. Introduction

Current Information Technologies and Web-based services need to manage, select and filter increasing amounts of textual information. Text classification allows users, through navigation on class hierarchies, to browse more easily the texts of their interests. This paradigm is very effective both in filtering information as in the development of online end-user services.

Since the number of documents involved in these applications is large, efficient and automatic approaches are necessary for classification. A Machine Learning approach can be used to automatically build the classifiers. The construction process can be seen as a problem of supervised learning: the algorithm receives a relatively small set of labelled documents and generates the classifier. Several algorithms have been applied, such as decision trees, linear discriminant analysis and logistic regression, the nave Bayes algorithm and Support Vector Machines (SVM). Besides having a justified learning theory describing its mechanics, SVM are known to be computationally efficient, robust and accurate.

Because of the globalization trend, an organization or individual often generates, acquires and archives the same document written in different lan-

guages (i.e., polylingual documents); moreover, many countries adopt multiple languages as their official languages. If these polylingual documents are organized into existing categories one would like to use this set of pre-classified documents as training documents to build models to classify newly arrived polylingual documents.

For multilingual text classification (i.e., collections of documents written in several languages), some prior studies address the challenge of cross-lingual text classification. However, prior research has not paid much attention to using polylingual documents yet. This study is motivated by the importance of providing polylingual text classification support to organizations and individuals in the increasingly globalized and multilingual environment.

We propose a method that combines different monolingual classifiers in order to get a new classifier as good as the best monolingual one which has the ability to deliver all the best performance measures (precision, recall and F1) possible.

This methodology was applied and evaluated on a set of legal documents from the EUR-Lex site. We collected documents for two anglo-saxon languages (English and German) and two roman ones (Italian and Portuguese), obtaining four different sets. The obtained results were quite good, indicating that combining different monolingual classifiers may be a promising approach to the problem of classifying documents written in several languages.

The paper is organized as follows: Section 2 describes the main concepts and tools used in our approach, Section 3 introduces the methodology for combining monolingual classifiers and Section 4 presents the document collection used for evaluation, describes the experimental setup and evaluates the obtained results. Finally, Section 5 presents some conclusions and points out possible future work.

2. Concepts and Tools

This section introduces the Automatic Text Classification approach and the classification algorithm and software tool used in this work.

2.1. Automatic Text Classification

Originally, research in Automatic Text Classification addressed the binary problem, where a document is either relevant or not w.r.t. a given category. However, in real-world situations the great variety of different sources and hence categories usually poses a multi-class classification problem, where a document belongs to exactly one category from a predefined set. Even more general is the multi-label problem, where a document can be classified into more than one category.

In order to be fed to the learning algorithm, documents must by preprocessed to obtain a more structured representation. The most common approach is to use a bag-of-words representation (Salton, 1975), where each document is represented by the words it contains, with their order and punctuation being ignored. Normally, words are weighted by some measure of word's frequency in the document and, possibly, the corpus. In most cases, a subset of words (stop-words) is not considered, because their role is related to the structural organization of the sentences and does not have discriminating power over different classes and some works reduce semantically related terms to the same root applying a lemmatizer.

Research interest in this field has been growing in the last years. Several machine learning algorithms were applied, such as decision trees (Tong, 1994), linear discriminant analysis and logistic regression (Schütze, 1995), the naïve Bayes algorithm (Mladenić, 1999) and Support Vector Machines (SVM)(Joachims, 1999). Joachims (Joachims, 2002) says that using SVMs to learn text classifiers is the first approach that is computationally efficient and performs well and robustly in practice. There is also a justified learning theory that describes its mechanics with respect to text classification.

2.1.1. *Multilingual text classification.*
While most text classification studies focus on monolingual documents, some point to multilingual text classification. From these, the great majority address the challenge of crosslingual text classification where the classification model relies on monolingual training documents and a translation mechanism to classify documents written in another language (Bel, 2003; Rigutini, 2005; Lee, 2009). A technique that takes into account all training documents of all languages when constructing a monolingual classifier for a specific language is proposed in (Wei, 2007). Wei et al. showed that for English and Chinese a feature-based reinforcement polylingual category integration approach obtains better accuracy then monolingual ones. Our proposal is quite different because we do not use information from other languages and multilingual thesaurus to build the individual classifiers. Our aim is to combine individual classifiers in order to obtain a better classifier and not to improve individual classifiers.

2.2. Support Vector Machines

Support Vector Machines, a learning algorithm introduced by Vapnik and coworkers (Cortes, 1995), was motivated by theoretical results from statistical learning theory: it joins a kernel technique with the structural risk minimization framework.

Kernel techniques comprise two parts: a module that performs a mapping from the original data space into a suitable feature space and a learning al-

gorithm designed to discover linear patterns in the (new) feature space. The *kernel function*, that implicitly performs the mapping, depends on the specific data type and domain knowledge of the particular data source.

The *learning algorithm* is general purpose and robust. It's also efficient since the amount of computational resources required is polynomial with the size and number of data items, even when the dimension of the embedding space grows exponentially (Shawe-Taylor, 2004). A mapping example is illustrated in Fig. 1a).

The *structural risk minimization* (SRM) framework creates a model with a minimized VC (Vapnik-Chervonenkis) dimension. This developed theory (Vapnik, 1998) shows that when the VC dimension of a model is low, the expected probability of error is low as well, which means good performance on unseen data (good generalization). In geometric terms, it can be seen as a search to find, between all decision surfaces (the $\mathcal{T}$-dimension surfaces that separate positive from negative examples) the one with maximum margin, that is, the one having a separating property that is invariant to the most wide translation of the surface. This property can be enlighten by Fig. 1b) that shows a 2-dimensional problem.

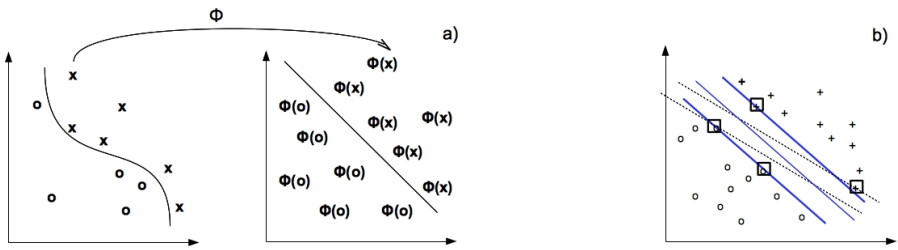

Figure 1. The SVM approach: kernel transformation and search for maximum margin.

2.2.1. *Classification software.*
As classification software we used SVM^{light} (Joachims, 1999)[1]. It is a C implementation of SVM that allows solving classification, regression and ranking problems, handles many thousands of support vectors and several hundred-thousands of training examples and supports standard kernel functions besides letting the user define its own.

3. Combining monolingual classifiers

Having documents in several languages, one can adopt a nave approach by considering the problem as multiple independent monolingual text classification problems. This simple approach only employs the training documents

[1] Available at `http://svmlight.joachims.org`

of one language to construct a monolingual classifier for that language and ignores all training documents of other languages. When a new document in a specific language arrives, one select the corresponding classifier to predict appropriate category(s) for the target document. However, the independent construction of each monolingual classifier fails to use the opportunity offered by polylingual training documents to improve the effectiveness of the classifier.

With this bearing in mind, and to get a decision for a new document, monolingual classifiers could be improved up in several ways. We propose the following strategies for the combination system:

- the sum of SVMs output values
- the F_1 weighted sum of SVMs output values
- the F_1 weighted sum of SVMs decisions

The above measures could also be used to draw decisions when considering a voting strategy of the monolingual classifiers.

4. Experiments

This section introduces the dataset, describes the experimental setup and presents the obtained results for the legal concepts classification task.

4.1. Dataset description

For testing the proposed methodology, experiments were run over a set of European Union law documents. These documents were obtained from the EUR-Lex site[2] within the "International Agreements" section, belonging to the "External Relations" subject matter. From all available agreements we chose the ones with full text (not just bibliographic notice) obtaining a set of 2714 documents (dated from 1953 to 2008).

Since agreements are available in several languages we collected them for two anglo-saxon languages (English and German) and two roman ones (Italian and Portuguese), obtaining four different corpora: `eurlex-EN`, `eurlex-DE`, `eurlex-IT` and `eurlex-PT`. Table I presents the total number and average per document of tokens (running words) and types (unique words).

Each document is classified onto several ontologies: the "EUROVOC descriptor", the "Directory code" and the "Subject matter". In all available classifications each document can be assigned to several categories. For our classification problem we used the first level of the "Directory code" classification, considering only categories with at least 50 documents. Table II shows each category along with the number of documents assigned.

[2] Available at `http://eur-lex.europa.eu/en/index.htm`

Table I. Total number and average per document of tokens and types for each corpus.

	tokens		types	
corpus	total	per doc	total	per doc
eurlex-EN	10699234	3942	73091	570
eurlex-DE	10145702	3728	133191	688
eurlex-IT	10665455	3929	96029	636
eurlex-PT	9731861	3585	86086	567

Table II. Number of documents assigned to each category.

id	*name*	*# of docs*
2	Customs Union and free movement of goods	209
3	Agriculture	390
4	Fisheries	361
7	Transport policy	81
11	External relations	2628
12	Energy	58
13	Industrial policy and internal market	55
15	Environment, consumers and health protection	138
16	Science, information, education and culture	99

4.2. EXPERIMENTAL SETUP

The experiments were done using a bag-of-words representation of documents, the SVM algorithm was run using SVM^{light} with a linear kernel and other default parameters and the model was evaluated using a 10-fold stratified cross-validation procedure with significance tests done with a 90% confidence level.

To represent each document we used the bag-of-words approach, a *vector space model* (VSM) representation where each document is represented by the words it contains, with their order and punctuation being ignored. Document's representation was obtained by mapping all numbers to the same token and using the tf-idf weighting function normalized to unit length.

To measure learner's performance we analyzed precision, recall and the F_1 measures (Salton, 1975) of the positive class. These measures are obtained from contingency table of the classification (prediction *vs.* manual classification).

4.3. MONOLINGUAL EXPERIMENTS

To support our claim, as baseline we have built classifiers for each language. Table III shows the average precision, recall and F_1 measures for each corpus and each category (boldface values are significantly worse than the best value obtained). Last line presents the average values over all nine classes.

Table III. Average precision, recall and F_1 values for each mono-lingual classifier.

	precision				*recall*				F_1			
id	EN	DE	IT	PT	EN	DE	IT	PT	EN	DE	IT	PT
2	**.919**	.957	**.922**	.937	.651	.665	**.580**	**.565**	.755	.778	**.702**	**.701**
3	.916	.928	.938	.943	.818	.805	**.705**	**.503**	.862	.860	**.803**	**.655**
4	**.956**	.966	.980	.971	.934	.906	.914	**.823**	.944	.934	.945	.890
7	.846	.870	**.793**	.806	.568	.543	.518	.482	.651	.640	.608	.590
11	.973	.973	.973	.973	.998	.997	.998	.997	.985	.985	.985	.985
12	.958	**.874**	**.877**	.938	.637	.700	.670	.600	.752	.765	.745	.716
13	.942	.933	.933	.967	.393	.320	.300	.320	.522	.454	.436	.461
15	.909	.922	.917	.908	.726	.732	.725	.732	.801	.813	.805	.806
16	**.862**	**.883**	.916	.947	.779	.799	.718	.647	.804	.828	.785	**.753**
avg	.828	.832	.825	.839	.650	.647	.613	.567	.708	.706	.681	.656

For the precision values we can notice that the Portuguese dataset has values with no significant difference with the "best" for all classes; all other languages perform worse for some classes (English: $c2$, $c4$ and $c16$; German: $c12$ and $c16$; Italian: $c2$, $c7$ and $c12$). With this in mind one can say that the Portuguese language generates the best precision classifiers.

Concerning recall, it's the English and German languages that consistently present the best values; Italian and Portuguese while equally good for some classes, are worse for others (Italian: $c2$ and $c3$; Portuguese: $c2$, $c3$ and $c4$).

The F_1 measure presents the same behavior as recall, being the only difference the classes where the Portuguese language performs worse ($c2$, $c3$ and $c16$).

4.4. POLYLINGUAL EXPERIMENTS

From all possible combiners (see Section 3), there is one that, for all classes, persistently generated the best F_1 values: the F_1 weighted sum of SVMs decisions.

Table IV shows, for each performance measure its results compared with the "best" monolingual classifiers(boldface values are significantly worse than the corresponding multilingual one): the Portuguese one for precision, and

the English and German one for recall and F_1. Last line equally presents the average values over all classes.

Table IV. Average precision, recall and F_1 values compared with the combiner ones.

	precision		*recall*			F_1		
id	PT	comb	EN	DE	comb	EN	DE	comb
2	.937	.947	.651	.665	.675	.755	.778	.782
3	.943	.925	.818	.805	.813	.862	.860	.863
4	.971	.964	.934	**.906**	.928	.944	.934	.945
7	.806	.868	.568	.543	.567	.651	.640	.654
11	.973	.973	.998	.997	.998	.985	.985	.985
12	.938	.908	.637	.700	.670	.752	.765	.761
13	.967	.933	.393	.320	.340	.522	.454	.467
15	.908	.912	.726	.732	.754	.801	.813	.821
16	.947	.881	.779	.799	.779	.804	.828	.815
avg	.839	.831	.650	.647	.652	.708	.706	.709

From the average values, one can easily see that precision is higher than recall and that the best monolingual classifier depends on what performance measure one is considering. Nevertheless, the combined classifier has all performance measures very similar and never significatively worse then the best monolingual classifier.

In fact, significant tests show that, for all classes and all performance measures, there is no significant difference between the "best" monolingual classifier and the corresponding combined classifier.

5. Conclusions and Future Work

A proposal to combine monolingual classifiers was presented and evaluated. The proposed methodology uses SVM classifiers to associate concepts to legal documents and uses a decision function that combines them in order to obtain, for each class, a classifier as good as the best monolingual classifier of each performance measure.

The baseline experiments allows one to conclude that some languages generate classifiers with better precision values (Portuguese language) while others generate classifiers with better recall ones (English and German languages). In order to be able to explain and to try to generalise these results further experiments need to be done. For instance, we will need to evaluate this methodology with other collections and domains. Are these results

specific for the legal domain? Or only for this collection and topics? Nevertheless, from a linguistic point of view, these results raise quite interesting questions.

By combining all classifiers one obtains a classifier as good as the best monolingual one. This combined classifier can even be considered better than the others since it has the ability to deliver all the best performance measures (precision, recall and F_1) unlike using one monolingual classifier.

As ongoing research we intend to use a deeper linguistic representation of documents and to re-evaluate this methodology. Specifically, we will use a semantic representation (based on DRS[3]) of documents and a graph kernel to create SVM models. In previous work, this approach showed to be able to improve the bag-of-words result for the Portuguese language. Another research line is to use legal thesaurus, such as the LOIS[4] lexical thesaurus, to reinforce some features/terms. With this approach we would combine our proposal with the main ideas of the Wei et al. work (Wei, 2007).

References

Bel, N., Koster, C. and Villegas, M. (2003), *Cross-lingual text categorization*, in Proceedings of ECDL'03, Proceedings of the 7th European Conference on Research and Advanced Tecnology for Digital Libraries, pp. 126–139.

Cortes, C. and Vapnik, V. (1995), *Support-vector networks*, Machine Learning, Vol. 20 No. 3, pp. 273–297.

Joachims, T. (1999a), *Making large-scale SVM learning practical*, in Schölkopf, B., Burges, C. and Smola, A. (Ed.), "Advances in Kernel Methods - Support Vector Learning", MIT Press.

Joachims, T. (2002), *Learning to Classify Text Using Support Vector Machines*, Kluwer Academic Publishers.

Lee, C.H. and Yang, H.C. (2009), *Construction of supervised and unsupervised learning systems for multilingual text categorization*, Expert Systems Applications, Vol. 36 No. 2, pp. 2400–2410.

Mladenić, D. and Grobelnik, M. (1999), *Feature selection for unbalanced class distribution and naïve Bayes*, in Proceedings of ICML'99, 16th International Conference on Machine Learning, pp. 258–267.

Rigutini, L., Maggini, M. and Liu, B. (2005), *An EM Based Training Algorithm for Cross-Language Text Categorization*, in Proceedings of WI'05, IEEE/WIC/ACM International Conference on Web Intelligence (IEEE Computer Society), pp. 529–535.

Salton, G., Wang, A. and Yang, C. (1975), *A vector space model for information retrieval*, Journal of the American Society for Information Retrieval, Vol. 18, pp. 613–620.

Schütze, H., Hull, D. and Pedersen, J. (1995), *A comparison of classifiers and document representations for the routing problem*, in Proceedings of SIGIR'95, 18th International Conference on Research and Developement in Information Retrieval (ACM), pp. 229–237.

Shawe-Taylor, J. and Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press.

[3] Discourse Representation Structures

[4] Lexical Ontologies for Legal Information Sharing

Tong, R. and Appelbaum, L.A. (1994), *Machine learning for knowledge-based document routing*, in Proceedings of TRC'94, 2nd Text Retrieval Conference.

Vapnik, V. (1998), *Statistical learning theory*, Wiley, NY.

Wei, C., Shi, H. and Yang, C. (2007), *Feature reinforcement approach to poly-lingual text categorization*, in Proceedings of the International Conference on Asia Digital Libraries (LNCS Springer), pp. 99–108.

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*.

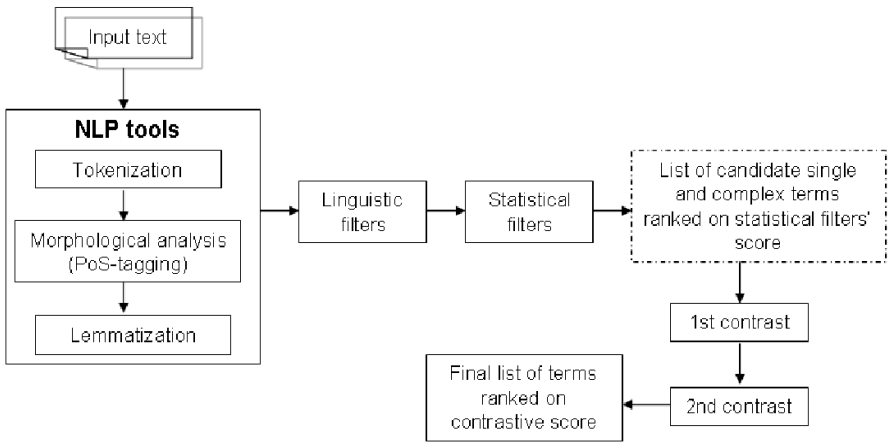

Figure 1. Term Extraction Process

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.

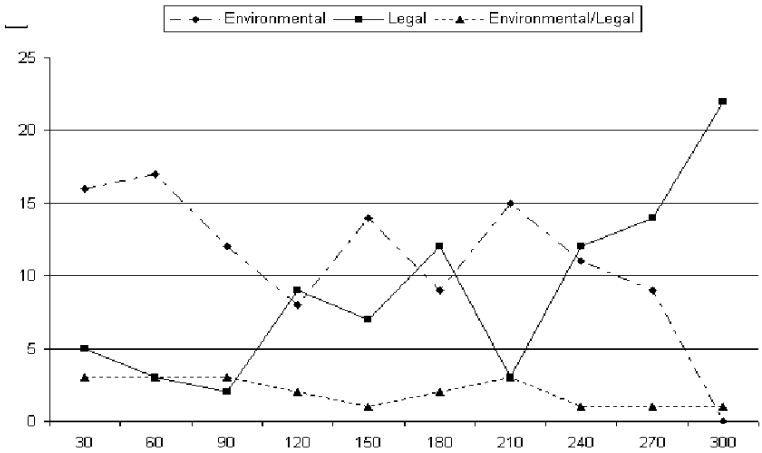

Figure 2. Distribution of the three types of terms in the extracted multi–word term list

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.

SECTION II

Legal Knowledge Modelling

A URN Standard for Legal Document Ontology: a Best Practice in the Italian Senate

Enrico Francesconi*, Carlo Marchetti°, Remigio Pietramala°, Pierluigi Spinosa*
*Institute of Legal Information Theory and Techniques of CNR (ITTIG-CNR), Florence, Italy
°Senate of the Republic, Italy

Abstract. Uniform Resource Names (URNs) are conceived by the Internet community for providing unambiguous and lasting identifiers of network resources, independently from their physical locations, availability and actual publication. In this paper a proposal of a URN schema for indentifying sources of law at international level is presented. Moreover an implementation of such schema at the Italian Senate is shown.

Keywords: Sources of law, Internet resources identification, URN

1. Introduction

Uniform Resource Names (URNs) are conceived by the Internet community for providing unambiguous and lasting identifiers of network resources, independently from their physical locations, availability and actual publication. In particular they play a key role in the legal domain where references to other legislative measures are very frequent and extremely important: the possibility of being able to immediately providing effective references and accessing legal documents is a desirable feature able to promote transparency and "certainty of law". Moreover the growing necessity of improved quality and accessibility of legal information amplifies the need for interoperability among legal information systems in national and international setting. A persistent, shared, open standard identifier for legal documents at international level is an essential prerequisite for establish such interoperability. Besides legal content providers, Internet content creators including publishers operating well outside the traditional arenas of legal publishing (news, technical documentation providers, etc.) can benefit by this standard because it facilitates the linking of legal documents and reduces the cost of maintaining documents that contain such references. This will result in a benefit for users as well, since they will enjoy a more richness and reliability of cross-referencing facilities, not only limited within the same information system as it is usually today. In the last few years a number of initiatives both in and outside Europe have arisen in the field of legal document standards to improve legal document accessibility on the Internet (Francesconi 2007). In this paper we describe a standard for the identification of sources of law,

recently submitted to the IETF as Internet Draft[1]: it is based on a URN technique capable of scaling beyond national boundaries as well as on the definition of a namespace convention (LEX) and a structure that will create and manage identifiers for sources of law at international level. The identifiers will be globally unique, transparent, persistent, location-independent, and language-neutral. These qualities will facilitate legal document management, moreover they will provide a mechanism of stable cross-collections and cross-country references. In this direction also the Permanent Bureau of the Hague Conference on Private International Law has recently expressed its opinion, encouraging EU Member States to adopt neutral methods of citation of their legal materials, including methods that are medium-neutral, provider-neutral and internationally consistent. This paper is organized as follows: in Section 2 the general structure of the URN-LEX identifier is introduced; in Section 3 the bibliographic FRBR reference model which the URN-LEX schema is based on is described; in Section 4, 5, 6 and 7 the main components of the schema able to identify legal documents at different levels of abstraction are shown; in Section 8 the modalities to establish references to a whole document or part of it using the URN-LEX methodology is briefly discussed; in Section 9 the principles of the resolution service are described; in Sections 10 and 11 the URN-LEX schema and a tool for automatic legal references mark-up according to such standard as implemented within the Italian Senate Web site are respectively described. Finally in Section 12 some conclusions are reported.

2. Structure of the identifier

As usual, the problem is to provide the right amount guidance at the core of the standard while providing sufficient flexibility to cover a wide variety of needs. The proposed URN- LEX identifier standard does this by splitting the identifier into a hierarchy of components. Its main structure is:

"urn:lex:"<NSS>

where “urn:lex” is the Namespace, which represents the domain in which the name has validity, as well as NSS is the Namespace Specific String composed as follows:

<NSS>::=<country>":"<local-name>

where: <country> is the part providing the identification of the country, or the multi-national or international organisation, issuing the source of law;

[1] http://datatracker.ietf.org/doc/draft-spinosa-urn-lex/

<local-name> is the uniform name of the source of law itself. It is able to represent all the aspects of an intellectual production, as it is a legal document, from its initial idea, through its evolution during the time, to its realisation by different means (paper, digital, etc.).

The <country> element is composed of two specific fields:

<country>::=<country-code>[";"<country-unit>]*

where: <country-code> is the identification code of the country where the source of law is issued. This code follows the standard [ISO 3166] Alpha-2 (it=Italy, fr=France, dk=Denmark, etc.). In case of multi-national (e.g., European Union) or international (e.g., United Nations) organizations the Top Level Domain Name (e.g., "eu") or the Domain Name (e.g., un.org, wto.int) is used instead of ISO 3166 code; <country-unit> are the possible administrative hierarchical sub-structures defined by each country, or organization, according to its own structure. This additional information can be used where two or more levels of legislative or judicial production exist (e.g., federal, state and municipality level) and the same bodies may be present in each jurisdiction. Then acts of the same type issued by similar authorities in different areas differ for the country-unit specification.

3. Reference Model for the <local-name> structure

The <local-name> will encode all the aspects of an intellectual production, from its initial idea, through its evolution during the time, to its realisation by different means (paper, digital, etc.). For these purposes it is based on the FRBR[2] model developed by IFLA[3]. Following the FRBR model, in a source of law, as in any intellectual production, 4 fundamental entities (or aspects) can be specified.

The first 2 entities reflect its contents: Work: identifies a distinct intellectual creation; in our case, it identifies a source of law both in its being (as it has been issued) and in its becoming (as it is modified over time); Expression: identifies a specific intellectual realisation of a work; in our case it identifies every different (original or up-to-date) version of the act over time and/or language in which the text is expressed;

while the other 2 entities relate to its form:

Manifestation: identifies a concrete realisation of an expression; in our case it identifies realizations in different media (printing, digital, etc.), encoding formats (XML, PDF, etc.), or other publishing characteristics; Item:

[2] Functional Requirements for Bibliographic Record

[3] International Federation of Library Associations and Institutions

identifies a specific copy of a manifestation; in our case it identifies individual physical copies as they are found in particular physical locations.

4. Structure of the <local-name>

The <local-name> component of the urn:lex identifier contains all the necessary pieces of information enabling the unequivocal identification of a legal document, within a specific legal system. In the urn:lex specification, a legal resource at "work" level is identified by four elements: the enacting authority; the type of measure; details (or terms) (like date of issue, number of the act, etc.) possibly, any annex.

It is often necessary to differentiate various expressions, that is: the original version and all the amended versions of the same document; the versions of the text expressed in the different official languages of the state or organization.

Finally the uniform name allows a distinction among diverse manifestations, which may be produced in multiple locations using different means and formats. In every case, the basic identifier of the source of law (work) remains the same, but information is added regarding the specific version under consideration (expression); similarly a suffix is added to the expression for representing the characteristics of the publication (manifestation). All this set of information is expressed in the jurisdiction official language; in case of more official languages, more names (aliases) are created for each language.

Therefore, the more general structure of the national name appears as follows:

```
<local-name>::=<work>[``@''<expression>]?["$"<manifestation>]?
```

However, consistent with legislative practice, the uniform name of the original provision becomes the identifier of an entire class of documents which includes: the original document, the annexes, and all its versions, languages and formats subsequently generated.

5. Structure of the Identifier at Work Level

The structure of the document identifier at work level is made of the four fundamental elements mentioned above, chosen from those used in citations, clearly distinguished one from the other in accordance with an order identifying increasingly narrow domains and competences. The use of citation elements at work level allows to construct the URN of the cited act manually or by software tools implementing automatic hyperlinking of legal sources on the basis of the textual citations of the acts. The general structure of the identifier at work level is:

```
<work>::=<authority>":"<measure>":"<details>[":"<annex>]*
```

where:
<authority> is the issuing authority of the measure (e.g., State, Ministry, Municipality, Court, etc.);
<measure> is the type of the measure (e.g., act, decree, decision, etc.);
<details> are the terms associated to the measure, typically the date and the number;
<annex> is the identifier of the annex, if any (e.g., Annex 1).

In case of annexes, both the main document and its annexes have their own uniform name so that they can individually be referenced; the identifier of the annex adds a suffix to that of the main document. In similar way the identifier of an annex of an annex adds an ending to that of the annex which it is attached to. The main elements of the national name are generally divided into several elementary components, and, for each, specific rules of representation are established (criteria, modalities, syntax and order)[4]. Examples of <work> identifiers are:

```
urn:lex:it:stato:legge:2006-05-14;22
urn:lex:uk:ministry.justice:decree:1999-10-07;45
urn:lex:ch;glarus:regiere:erlass:2007-10-15;963
urn:lex:es:tribunal.supremo:decision:2001-09-28;68
```

In the states or organisations that have more than one official language, a document has more identifiers, each of them expressed in a different official language, basically a set of equivalent aliases. This system permits manual or automated construction of the uniform name of the referred source of law in the same language used in the document itself (e.g., `urn:lex:eu:council:directive:2004-12-07;31`, `urn:lex:eu:consiglio:direttiva:2004-12-07;31`, etc.). Moreover, a document can be assigned more than one uniform name in order to facilitate its linking to other documents. This option can be used for documents that, although unique, are commonly referenced from different perspectives. For example, the form of a document's promulgation and its specific content (e.g., a Regulation promulgated through a Decree of the President of the Republic).

6. Structure of the Identifier at Expression Level

There may be several expressions of a legal text, connected to specific versions or languages. Each version is characterized by the period of time during which that text is to be considered as the valid text (in force or effective). The lifetime of a version ends with the issuing of the subsequent version. New

[4] For the details regarding each element, see Attachment B of the IETF Internet Draft http://datatracker.ietf.org/doc/draft-spinosa-urn-lex/

versions of a text may be brought into existence by: changes as regards text or time (amendments) due to the issuing of other legal acts and to the subsequent production of updated or consolidated texts; correction of publication errors (rectification or errata corrige); entry into or departure from a particular time span, depending on the specific date in which different partitions of a text come into force. Each such version may be expressed in more than one language, with each language-version having its own specific identifier. The identifier of a source of law expression adds such information to the work identifier, using the following main structure:

```
<expression>::="@"<version>[":"<language>]?
```

where:
`<version>` is the identifier of the version of the (original or amended) source of law. In general it is expressed by the promulgation date of the amending act; anyway other specific information can be used for particular cases. If necessary, the original version is specified by the string "original";
`<language>` is the identification code of the language in which the document is expressed, according to ISO 639-1 [7] (it=Italian, fr=French, de=German, etc.); in case the code of a language is not included in this standard, the ISO 639-2 (3 letters) is used. This information is not necessary when the text is expressed in the unique official language of the country.

Examples of document identifiers for expressions are:
`urn:lex:ch:etat:lois:2006-05-14;22@originel:fr` (original version in French)
`urn:lex:ch:staat:gesetz:2006-05-14;22@original:de` (original version in German)
`urn:lex:ch:etat:lois:2006-05-14;22@2008-03-12:fr` (amended version in French)
`urn:lex:ch:staat:gesetz:2006-05-14;22@2008-03-12:de` (amended version in German)

7. Structure of the Identifier at Manifestation Level

To identify a specific manifestation, the uniform name of the expression is followed by a suitable suffix describing the: digital format (e.g., XML, HTML, PDF, etc.) expressed according to the MIME Content-Type standard [RFC 2045], where the "/" character is to be substituted by the "-" sign; publisher or editorial staff who produced it; possible components of the expressions contained in the manifestation. Such components are expressed by "body" (the default value), representing the whole or the main part of the document, or by the caption of the component itself (e.g. Table 1, Figure 2, etc.); other features of the document (e.g., anonymized decision text).

The `<manifestation>` suffix will thus read:

```
<manifestation>::=<format>":"<editor>[":"<component>]?[":"<feature>]?
```

To indicate possible features or peculiarities, each principal element of the manifestation may be followed by a further specification. For example, the original version the Italian act 3 April 2000, n. 56 might have the following manifestations with their relative uniform names:

PDF format (vers. 1.7) of the whole act edited by the Parliament:
`urn:lex:it:stato:legge:2000-04-03;56$application-pdf;1.7:parliament`

Furthermore, it is useful to be able to assign a uniform name to a component of a manifestation in case non-textual objects are involved. These may be multimedia objects that are non-textual in their own right (e.g. geographic maps, photographs, etc.), mixed with textual parts. In these ways, a "lex" name permits: exploitation of all the advantages of an unequivocal identifier that is independent of physical location; a means to provide choice among different existing manifestations (e.g. XML or PDF formats, resolution degree of an image etc.) of the same expression.

8. Sources of Law References

References to sources of law often refer to specific partitions of the act (article, paragraph, etc.) and not to the entire document. Therefore, for allowing applications to manage this information(e.g., pointing a specific partition on the browser), it is necessary that a partition identifier within the act is present (i.e. an unequivocal label or ID). For enabling the construction of the partition identifier between different collections of documents, specific construction rules for IDs or labels SHOULD be defined and shared, within each country or jurisdiction, for any document type (e.g., for legislation, the paragraph 2 of the article 3 might have as label or ID the value "art3-par2").

Furthermore, it is useful to foresee the compatibility with applications able to manage this information (e.g., returning the proper element); these procedures are particularly useful in the case of rather long acts, such as codes, constitutions, regulations, etc.

For this purpose it is necessary that the partition identifier is transmitted to the servers (resolution and application) and therefore it cannot be separated by the typical "#" character of URI fragment, which is not transmitted to the server.

According to these requirements, the syntax of a reference is:

```
<URN-reference>::=<URN-document>["~"<partition-id>]?
```

(e.g., to refer to the paragraph 3 of the article 15 of the French Act of 15 may 2004, n. 106, the reference is written
`urn:lex:fr:etat:loi:2004-05-15;106~art15-par3`).

Using a different separator ("~") from the document name, the partition ID is not withheld by the browser but it is transmitted to the resolution process. This enables the resolver to retrieve (for example, out of a database), if it is possible, only the referred partition, otherwise to return the whole act. Anyway, to make it effective pointing to the indicated partition, the resolver SHOULD transform the partition ID of each returned URL in a URI fragment; this is obtained appending to URL the "#" character followed by the partition ID (in the example above, the returned URL will be <URL-document>#art15-par3).

Anyway it is possible to use the general syntax (with "#"); in this case only the URN document component of the reference is transmitted to the resolver, therefore the whole document will be always retrieved.

9. The Resolution Service

The task of the resolution service is that of associating a LEX identifier with a specific document address on the network. The system has a distributed architecture based on two fundamental components: a chain of information in DNS (Domain Name System) and a series of resolution services from URNs to URLs, each competent within a specific domain of the namespace. Through the NAPTR records of the DNS (described in [RFC 3403]), the client identifies the characteristics (protocol, port, site) of the service capable of associating the relative URLs with the URN in question, thereby allowing access to the document. A resolution service can delegate the resolution and management of hierarchically-dependent portions of the name. Delegation of this responsibility will not be unreasonably withheld provided that the processes for their resolution and management are robust and are followed. For the "lex" namespace, the declared registrant of the namespace (ITTIG-CNR) will maintain the root zone "lex.urn.arpa" and, in correspondence with the adhesion of a new country (e.g., "br"), will update the DNS information with a new record to delegate the relative resolution. This may be obtained by a regular expression that matches the initial part of the URN (e.g., "urn:lex:br") and redirects towards the proper zone (e.g., "lex.senado.gov.br"). Likewise the institution responsible for the country uniform names (e.g., "urn:lex:br") has the task of managing the relative root in the DNS system (e.g., "lex.senado.gov.br" zone) and routing the resolution towards its resolvers on the basis of parts of the uniform names. In similar way it can delegate the resolution of country sub-levels (e.g., "urn:lex:br;sao.paolo") towards the relative zone (e.g., "lex.sao-paolo.gov.br"). At the end of the delegation chain routing, the address of the resolution service is provided and this service gives back the network addresses (URLs) of the items. The resolution service is based on two main elements: a knowledge base (consist-

ing in a catalogue or a set of transformation rules) and a software to query the knowledge base itself.

9.1. Catalogues for Resolution

The architecture of the catalogue of resolution has to take into account that incompleteness and inaccuracy are rather frequent in legal citations, and incomplete or inaccurate uniform names of the referred document are thus likely to be built from textual references (this is even more frequent if they are created automatically through a specific parser). By contrast with systems that can be constructed around rigorous and enforceable engineering premises, such as DNS, the LEX resolver will be expected to cope with a wide variety of "dirty" inputs, particularly those created by the automated extraction of references from incomplete or inaccurate texts. In this document, the result is a particular emphasis on a flexible and robust resolver design. For these reasons, the implementation of a catalogue, based on a relational-database, is suggested, as it will lead to a more higher flexibility in the resolution process as partial match. In addition the catalogue must manage the aliases, the various versions and languages of the same source of law as well as the related manifestations. It is suggested that each enacting authority implements its own catalogue, assigning a corresponding unambiguous uniform name to each resource.

9.2. Suggested Resolver Behaviour

First of all the resolution process should implement a normalization of the uniform name to be resolved. This may involve transforming some components to the canonical form (e.g., filling out the acronyms, expanding the abbreviations, unifying the institution names, standardizing the type of measures, etc.). For this function the registers of names and authorities organization, including validity time span, as well as the registers of the types of measure are useful. The resolver should then query the catalogue searching for the URN which corresponds exactly to the given one (normalized if necessary). Since the names coming from the references may be inaccurate or incomplete, an iterative, heuristic approach (based on partial matches) is suggested. It is worth remarking that incomplete references (not including all the elements to create the canonical uniform name) are normal and natural; for a human reader, the reference would be "completed" by contextual understanding given by the including document. Lacking more specific indications, the resolver should select the best (most recent) version of the requested source of law, and provide all the manifestations with their related items. A more specific indication in the uniform name to be resolved will, of course, result in a more selective retrieval, based on any suggested expression and/or manifestations components (e.g. date, language, format, etc.).

10. URN standard within the Italian Senate

URN:LEX standard has stemmed from the experience of the Italian legislative XML project NormeInRete (NIR). The feasibility study of such a project was launched in 1999, while the real implementation of the system started in 2001. A URN naming convention for legal resources was in particular defined, in terms of a URN:NIR namespace, whose structure shares, with the URN:LEX standard, principles, characteristics and identification components, therefore it can be considered an ante-litteram implementation of the URN:LEX naming convention. Due to these relationships a change from the NIR to the LEX more general namespace is straightforward and can be automatically implemented. Currently within the Italian Senate of the Republic Web site, a URN:NIR standard is implemented to identify the following type of documents: Assembly reports, Assembly agenda, Committee reports and minutes, Bills, Bill relations, Bill preambles, "Iter Legis" cards, Questions and answers reports. A transparent identifier for the previously mentioned types of documents are constructed, starting from the formal parameters of the acts. Here below are some examples:

Assembly report n. 365 of the XVI Legislature
`urn:nir:senato.repubblica;assemblea:resoconto:16.legislatura;365`

Assembly agenda of 15 April 2010
`urn:nir:senato.repubblica;assemblea:ordine.giorno:2010-04-15`

Committee report n. 259 of the XVI Legislature
`urn:nir:senato.repubblica;commissioni:bollettino:16.legislatura;259`

Bill n.1880 of the XVI Legislature
`urn:nir:senato.repubblica:disegno.legge:16.legislatura;1880`

Relation (template A) to the Bill n. 1880 of the XVI Legislature
`urn:nir:senato.repubblica:disegno.legge;relazione:16.legislatura;1880-a`

Approved preamble to the Bill n. 1880 of the XVI Legislature
`urn:nir:senato.repubblica:disegno.legge;approvato:16.legislatura;1880`

Iter Legis card between chambers, n. 1880 of the XVI Legislature
`urn:senato-it:parl:ddl:senato;16.legislatura;1880`

11. A tool for automatic legal references mark-up within the Italian Senate Web site

A legal text may contain lots of references to other documents which are described using the related URN, so that references can be transformed in effective links when documents are published on the Web. Information for URN construction is usually contained in citations (for example the citation: "Act 24 November 1999, No. 468" generates the following URN-NIR `urn:`

`nir:stato:legge:1999-11-24;468`). The manual construction of hyperlinks in terms of URN for each reference can be a time-consuming work. For this reason a module able to automatically parse legal documents, detecting cross-references and assigning them the related URNs has been developed. Such module, called xmLegesLinker, developed by ITTIG-CNR under the GNU-GPL license, is generated using LEX and YACC technologies (Johnson, 1975; Lesk, 1975), on the basis of the vocabulary of the citations and the URN grammar expressed in EBNF syntax.

Figure 1. LEX technologies

Using LEX technologies a lexical analyzer is generated (yylex) able to detect tokens, namely symbols (words, numbers and punctation marks) belonging to the citation vocabulary (Figure 1). Then using YACC technologies, a syntactical analyzer is generated (yyparse) able to recognize a sequence of tokens, generated by LEX, as representing a reference, and to construct the related URN (Figure 2).

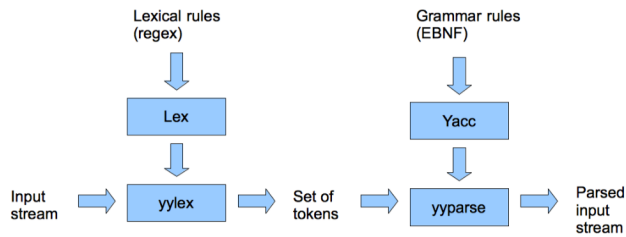

Figure 2. Combination of LEX and YACC technologies

Such tool is integrated within xmLegesEditor[5] a legislative XML editor developed by ITTIG-CNR for the NIR project, and it is used by several projects using NIR standards. In particular xmLegesLinker has been integrated within the Italian Senate Web site: once a document is queried through the Senate search engine, retrieved and displayed in the browser, the user may decide to automatically detect all the legal references in the text, as well as construct and display them, ready to query the Senate resolution system. For instance, given a citation to "Article 14 of Act 23 August 1988, No. 400", such reference is automatically detected and described according to the related URN: `urn:nir:stato:legge:1988-08-23;400~art14`

[5] http://www.xmleges.org

Moreover, such a URN is made effective by constructing a query to the Senate resolution system: `http://www.senato.it/uri-res/N2Ls?urn:nir:stato:legge:1988-08-23;400~art14` The resolution system will translate the URN into an automatic query addressed to a professional and commercial legislative database, in case the user is directly connected to the Senate intranet structure; otherwise, in case of internet users, the query will be automatically addressed to the public legislative database. Figure 3 shows a document retrieved within the Senate Web site, before and after the activation of the automatic references mark-up service (xmLegesLinker). The Senate resolution system makes it also possible to translate URN references to official internal publications, such as, to give an example: `http://www.senato.it/uri-res/N2Ls?urn:nir:senato.repubblica;assemblea:resoconto:16.legislatura;365`

Figure 3. Document before and after legal references mark-up

As far as internal users are concerned (Intranet users), the Senate made it available two further functions for the parsing of legal references:

1. Parsing of personal documents in the following formats: “plain text”, HTML, RTF, MS Word
2. Parsing of Internet sites.

The parsing of the users personal documents can only be made from computers within the Senate net. The following image shows the starting screenshot of the application:

In order to activate the function for the parsing of legal references, users must select a file type “plain text", HTML, RTF or MS Word from the file system.

Therefore, the application will show an HTML page consisting of two columns. The left column shows the original document in HTML format, whose legal references identified by the parser are highlighted. In case of activation of one of the links in the left column, the right column shows the result of the search, that is to say, the text of the legal resource retrieved in the professional legislative database used by the Senate.

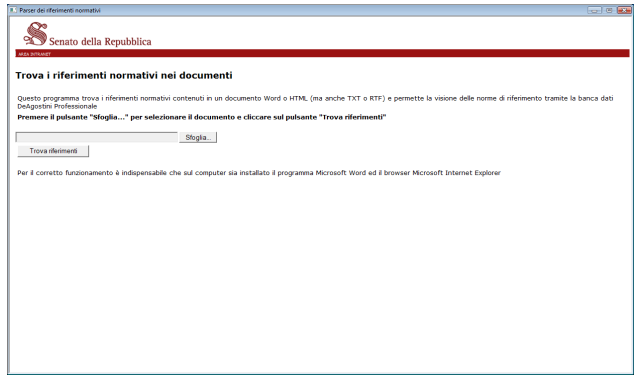

Figure 4. Parsing of personal documents – Start page

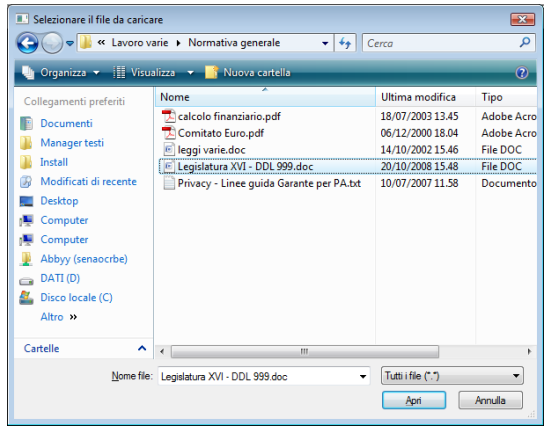

Figure 5. Parsing of personal documents – Document choice

The application is based on the integrated use of an MS Word converter (whose presence in the user's computer is mandatory), of the parser xmLeges-Linker and of the Senate URN2DEA resolver, which automatically translates a URN:NIR identifier in a query addressed to the professional database used by the Senate. The second parsing function, available only for internal users of the Senate net, enables the scanning of legal references which may occur in any internet site. Users only need to enter a page URL into the starting page "Find Legal References".

Clicking on the "Find" button, the original webpage is captured and parsed, then the detected legal references are highlighted with a link. Basically, this service occurs between the browser and the requested site (web proxy function); for each page, such service implements the references parsing by using xmLegesLinker. Similarly in this case the activation of a link invokes the URN2DEA Senate resolver. The following images show a legal Internet site, before and after the use of the above mentioned function:

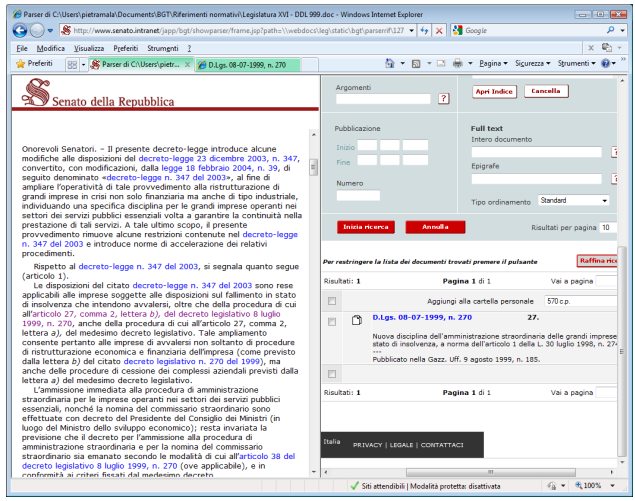

Figure 6. Parsing of personal documents – Result

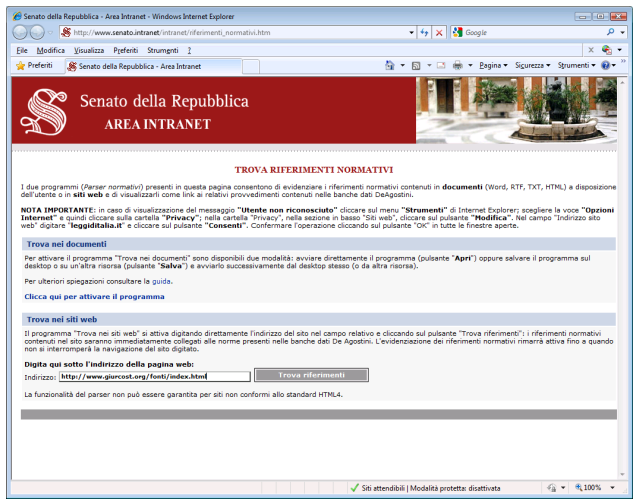

Figure 7. Parsing of Internet sites – Start page

12. Conclusions and Future Perspectives

In this paper the main principles of a URN schema for legal documents (sources of law) as submitted to IETF for registration in terms of a LEX namespace is presented. The syntax of the identifier and its usage in a multilanguage context is shown, as well as the principles of a resolution service able to guarantee persistence of the links based on URN, independently from

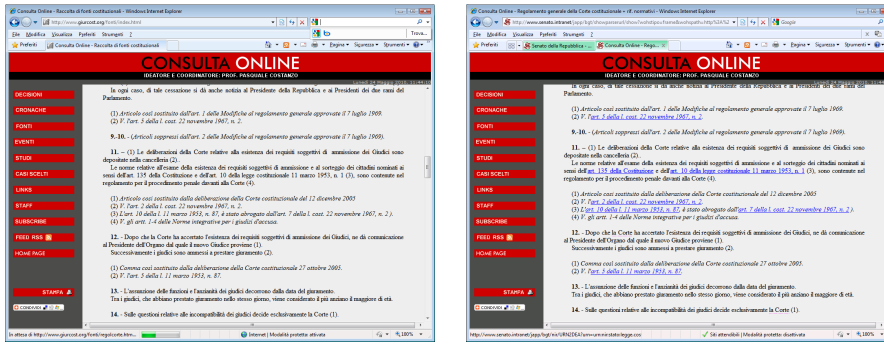

Figure 8. Parsing of Internet sites – Original and parsed web page

any change in document physical locations. The URN:LEX RFC is currently at the status of IETF Internet Draft and it is going to be revised according to the comments which are being received. Moreover an implementation of the URN:LEX standard within the Italian Senate of the Republic, as well as a tool to implement automatic legal references mark-up (automatic legal documents hyperlinking) as integrated within the Italian Senate Web site, have been shown. Shortly a plug-in for Firefox, developed by ITTIG-CNR, will be available: it allows a browser to natively exploit the URN protocol, routing the resolution service through the DNS Internet infrastructure, without the necessity to transform a URN hyperlink attribute into an http query to the resolution system.

References

P. Spinosa, E. Francesconi, C. Lupo, “A Uniform Resource Name (URN) Namespace for Sources of Law (LEX)”, May 2010, http://datatracker.ietf.org/doc/draft-spinosa-urn-lex/

S. Bradner, “Key words for use in RFCs to Indicate Requirement Levels”, BCP 14, RFC 2119, March 1997.

D Daigle, L., van Gulik, D., Iannella, R., and P. Faltstrom, “Uniform Resource Names (URN) Namespace Definition Mechanisms”, BCP 66, RFC 3406, October 2002.

R. Moats, K. R. Sollins, “URN Syntax”, RFC 2141, May 1997.

Berners-Lee, T., Fielding, R., and L. Masinter, “Uniform Resource Identifiers (URI): Generic Syntax”, STD 66, RFC 3986, January 2005.

M. Mealling, Dynamic Delegation Discovery System (DDDS), Part Three: The Domain Name System (DNS) Database, RFC 3403, October 2002.

Narten, T. and H. Alvestrand, “Guidelines for Writing an IANA Considerations Section in RFCs”, BCP 26, RFC 2434, October 1998.

ISO 3166, “Country name codes”, ISO 3166-1:1997.

ISO 639, “Codes for the representation of names of languages” - Part 1: alpha-2 code - Part 2: alpha-3 code. (1998, 2002)

R. Daniel, “A Trivial Convention for using HTTP in URN”, RFC 2169, June 1997

N. Freed, N. Borenstein, “Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies”, RFC 2045, November 1996.

P.L. Spinosa, “The Assignment of Uniform Names to Italian Legal Documents”, May, 2006

E. Francesconi, “Technologies for European Integration. Standards-based Interoperability of Legal Information Systems”, ISBN 978-88-8398-050-3, European Press Academic Publishing, 2007.

S.C. Johnson. Yacc - yet another compiler compiler. Technical Report CSTR 32, Bell Laboritories, Murray Hill, N.J., 1975.

M.E. Lesk. Lex - a lexical analyzer generator. Technical Report CSTR 39, Bell Laboritories, Murray Hill, N.J., 1975.

Using Intuitionistic Logic as a basis for Legal Ontologies

Edward Hermann Haeusler*, Valeria de Paiva◊ and Alexandre Rademaker°
*Depto de Inf. – PUC-Rio – Brasil
◊Curl Inc. – USA
°EPGE-FGV – Brasil

Abstract. Classical Description Logic has been widely used as a basis for ontology creation and reasoning in many knowledge specific domains. These specific domains naturally include Legal AI. As in any other domain, consistency is an important issue for legal ontologies. However, due to its inherently normative feature, coherence (consistency) in legal ontologies is more subtle than in most other domains. Negation and subsumption play a central role in ontology coherence. An adequate intuitionistic semantics for negation in a legal domain comes to the fore when we take legally valid individual statements as the inhabitants of our legal ontology. This allows us to elegantly deal with particular situations of legal coherence, such as conflict of laws, as those solved by Private International Law analysis. This paper: (1) Briefly presents our version of Intuitionistic Description Logic, called IALC for Intuitionistic ALC (ALC being the canonical classical description logic system)(2) Discuss the jurisprudence foundation of our system, and (3) Shows how we can perform a coherence analysis of "Conflict of Laws in Space" by means of IALC. This paper reports work-in-progress on using this alternative definition of logical negation for building and testing legal ontologies and reasoning in AI.

Keywords: Description logic, intuitionistic Logic, legal ontologies, constructive negation

1. Introduction

Classical Description Logic has been widely used as a basis for ontology creation and reasoning in many knowledge specific domains. These specific domains naturally include Legal AI. As in any other domain, consistency is an important issue for legal ontologies. However, due to its inherently normative feature, coherence (consistency) in legal ontologies is more subtle than in most other domains. Negation and subsumption play a central role in ontology coherence. An adequate intuitionistic semantics for negation in a legal domain comes to the fore when we take legally valid individual statements as the inhabitants of our legal ontology. This allows us to elegantly deal with particular situations of legal coherence, such as conflict of laws, as those solved by Private International Law analysis. This paper: (1) Briefly presents our version of Intuitionistic Description Logic, called IALC for Intuitionistic ALC (ALC being the canonical classical description logic system)(2) Discuss the jurisprudence foundation of our system, and (3) Shows how we can perform a coherence analysis of "Conflict of Laws in Space" by means of IALC. This paper reports work-in-progress on using this alternative definition of logical negation for building and testing legal ontologies and reasoning in AI.

2. A brief discussion on Jurisprudence and Intuitionism

One of the main problems from jurisprudence (legal theory) is to make precise the use of the term "law". In fact, the problem of individuation, namely, what counts as the unit of law, seems to be one of the fundamental open question in jurisprudence.Any approach to law classification requires firstly answering the question "What is to count as one complete law?" (Raz1972). There are two main approaches to this question.

One is to take the all (existing) legally valid statements as a whole. This totality is called "the law". This approach is predominant in legal philosophy and jurisprudence debiting his significance to the Legal Positivism tradition initiated by Hans Kelsen (for a contemporary reference see (Kelsen1991)). The coherence of "the law" plays a central role in this approach, whilst a debate whether coherence is built-in by the restrictions induced by Nature in an evolutionary way, or whether it should be object of knowledge management, seems to be a long and classical debate.

The other approach to law definition is to take into account all legally valid statements as being *individual laws*. This view, in essence, is harder to be shared with jurisprudence principles, since they firstly are concerned to justify law. This latter approach seems to be more suitable to Legal AI. It is also considered by Legal theoreticians, at least partially, whenever they start considering ontological commitments, such as, taking some legal relations as primitive ones (Hohfeld, 1919), *primary and secondary rule* (Hart, 1961) or even a two-level logic to deal with different aspects of law (see *logic-of-imperation/logic-of-obligation* from Bentham, 1970). In fact, some Knowledge Engineering (KE) groups pursue this approach as a basis for defining legal ontologies. We also follow this route. It is important to note that the pure use of a deontic logic has been shown to be inadequate to accomplish this task. In (Valente1995) it is shown that deontic logic does not properly distinguish between the normative status of a situation from the normative status of a norm (rule).

From the semantic point of view, iALC seems to be well suited to model the Legal theoretic approach pursued by KE as cited above. Let us consider an iALC model having as individuals each of the valid and possible *legal statements*. The $\preceq$ relation is the natural hierarchy existing between these individual *legal statements*, as well from any precedence relation related to them. For example, sometimes conflicts between *legal statements* are solved by inspecting the age of the laws (how old is the date of its first edition in the legal system), the wideness enforcement scope of each law, and etc. Any of these considered relations are order reactions. For example, "Theodor is vicariously liable by John" is legally dominated (precedes) by "John is a *worker* of Theodor", or "John and Theodor have an '*employment contract*'. Any legal statement involving the *civil* liability of someone must precedes

any legal statement asserting that he/she is of *legal age*. If C is a concept symbol in a description logic language, its semantics is the subset of *legal statements* representing a *kind of* legal situation.

The main role of the Intuitionism in our setting is the meaning it provides to the negation of concepts, as well as to subsumption.

Let us analyze briefly the case of negation of concepts regarding the classical ALC logic and a more traditional approach, based on classic ALC, to the ontological formalization of "the law" that includes the development of one or more domain ontologies to be used in validating the legal statements.

Consider a person, $Peter$, that is under the legal age in his living country. Let us consider, as it is usual, that people under the legal age are not able to sign contracts. Consider a part of an (hypothetical) ontology of *Private Ownership Law* with concepts $RentingContract$, for the set of valid renting contracts, $\exists hasTenant.RentingContract$, for the set of legal tenants, and, $\exists hasLandlord.RentingContract$ for the set of legal landlords. Of course, our $Peter$ is not in $\exists hasTenant.RentingContract$ nor in $\exists hasLandlord.RentingContract$ either. Thus, in classical ALC, $Peter$ is in the complement of each respective concept, namely $\neg\exists hasLandlord.RentingContract$ and $\exists hasLandlord.RentingContract$. In other words, "$Peter$ has no contract signed" has to be taken as a legal statement in our ontology. But, $Peter$ is under legal age, and hence, there must be no legal statement about him as an individual agent.

As seen in the previous paragraph, Classical negation forces the negation of a proposition to be part of a concept, but in the context of "the law" the negation of a valid law does not have to be valid either. Besides the ontological complexity of dealing with legal statements together with non-legal ones by defining concepts that are outside jurisprudence, Classical negation can lead to unnecessary incoherent situations in a legal ontology. The following paragraph illustrates this.

Suppose that $Peter$, from the above discussion, is under legal age in the place he lives, but he is citizen from another country where he is of legal age. If the country he lives has Private International Law then he has to be considered of legal age in the country he lives. Classical negation cannot be applied to this situation without leading to an incoherence: "Peter is and not is of legal age". The usual way to circumvent this kind of situation is to consider an auxiliary ontology on objects and agents outside "the law" but related to it, and, by means of these auxiliary terminological entities overcome the incoherency. A partial description of the world is then used to separate concerns in a way that the propositions cannot be seen as contradictory. For example, one may consider the auxiliary concept of $Foreigners$ with $Peter$ belonging to it because he is of legal age in his country. Of course, this solves the problem with $Peter$, but does not solve the problem with his children that are not of legal age in his country either.

From what we have discussed, we can conclude that in order to define a legal ontology, one has either to deal with parcels of the world that have to do more with application of the law than "the law" itself or to consider a different negation and propositions denoting valid aspects of "the law". The iALC approach depicted in this paper basically consists of a model which includes all the possible valid legal statements and the relations among them and the use of intuitionistic negation and subsumption instead of their classical counterparts.

3. Intuitionistic Description Logic iALC

Description logics are quite popular right now. However, They are classically biased, in the sense that the negation ($\neg$) of a concept is simply its set-theoretical complement, regarded to the universe of individuals. Subsumption of concepts is set-theoretical inclusion. This seems to be enough to most of the known applications. However, as discussed in (dePaiva2003), *constructive* description logics also makes sense, both from a theoretical and from a practical viewpoint. There are many ways of defining *constructive* description logics. In particular Mendler and Scheele have worked out an interesting system ((MS2008)). They cite auditing of business as their preferred application. Aiming to provide a formal basis for legal AI, we follow a different path and describe a constructive version of ALC, based on the framework for constructive modal logics developed by Simpson in his PhD thesis (Simpson1995). This framework was firstly developed by Brauner and de Paiva in (BdeP2006) for Hybrid Logics.

iALC is a basic description language. Its concept formers are described by the following grammar:

$$C, D ::= A \mid \bot \mid \top \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid C \sqsubseteq D \mid \exists R.C \mid \forall R.C$$

where A stands for an atomic concept and R for an atomic role. This syntax is more general than standard ALC in that it includes subsumption $\sqsubseteq$ as a concept-forming operator. Negation can be represented via subsumption, $\neg C = C \sqsubseteq \bot$, but we find it convenient to keep it in the language. The constant $\top$ can also be omitted since it can be represented by $\neg\bot$.

Following Mendler and Scheele we say a constructive interpretation of iALC is a structure $\mathcal{I} = (\Delta^{\mathcal{I}}, \preceq^{\mathcal{I}}, \cdot^{\mathcal{I}})$ consisting of a non-empty set $\Delta^{\mathcal{I}}$ of entities in which each entity represents a partially defined individual; a refinement preordering $\preceq^{\mathcal{I}}$ on $\Delta^{\mathcal{I}}$, i.e., a reflexive and transitive relation; and an interpretation function $\cdot^{\mathcal{I}}$ mapping each role name R to a binary relation $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ and each atomic concept A to a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ which is closed under refinement, i.e., $x \in A^{\mathcal{I}}$ and $x \preceq^{\mathcal{I}} y$ implies $y \in A^{\mathcal{I}}$. The

interpretation $\mathcal{I}$ is lifted from atomic $\bot, A$ to arbitrary concepts via:

$$
\begin{array}{rcl}
\top^{\mathcal{I}} & =_{df} & \Delta^{\mathcal{I}} \\
(\neg C)^{\mathcal{I}} & =_{df} & \{x | \forall y \in \Delta^{\mathcal{I}}.x \preceq y \Rightarrow y \notin C^{\mathcal{I}}\} \\
(C \sqcap D)^{\mathcal{I}} & =_{df} & C^{\mathcal{I}} \cap D^{\mathcal{I}} \\
(C \sqcup D)^{\mathcal{I}} & =_{df} & C^{\mathcal{I}} \cup D^{\mathcal{I}} \\
(C \sqsubseteq D)^{\mathcal{I}} & =_{df} & \{x | \forall y \in \Delta^{\mathcal{I}}.(x \preceq y \text{ and } y \in C^{\mathcal{I}}) \Rightarrow y \in D^{\mathcal{I}}\} \\
(\exists R.C)^{\mathcal{I}} & =_{df} & \{x | \forall y \in \Delta^{\mathcal{I}}.x \preceq y \Rightarrow \exists z \in \Delta^{\mathcal{I}}.(y,z) \in R^{\mathcal{I}} \text{ and } z \in C^{\mathcal{I}}\} \\
(\forall R.C)^{\mathcal{I}} & =_{df} & \{x | \forall y \in \Delta^{\mathcal{I}}.x \preceq y \Rightarrow \forall z \in \Delta^{\mathcal{I}}.(y,z) \in R^{\mathcal{I}} \Rightarrow z \in C^{\mathcal{I}}\}
\end{array}
$$

Clearly our setting is a simplification of Mendler and Scheele's where we dispense with infallible entities, since our system iALC satisfies (like classical ALC) $\exists R.\bot = \bot$. But $\exists R.(C \sqcup D) = \exists R.C \sqcup \exists R.D$, like in Mendler and Scheele's work is not necessarily true. We will have no use for nested subsumptions, but they do make the system easier to define, so we keep the general rules.

4. Applications of iALC

In this section we show an application of iALC to a part of Legal Ontology

We remind the reader that a concept symbol C, in a description logic language, is associated to a subset of *legal statements* representing a *kind of* legal situation. Roles in the description logic language are associated to relations between these *legal situations*, imposed by the relationship between each pair of individual *legal statements*.

In the sequel we detail the legal situation known as "Conflict of Laws in Space" within *Private International Law* scope. If ALCis used instead of iALC, the formal treatment is rather cumbersome. This is briefly commented at the end of the section.

Consider the following situation:

> Peter and Maria signed a renting contract. The subject of the contract is an apartment in Rio de Janeiro. The contract states that any dispute will go to court in Rio de Janeiro. Peter is 17 and Maria is 20. Peter lives in Edinburgh and Maria lives in Rio.

In order to exist in our model, the legal statement (1) *Maria and Peter have contractual obligations and rights to each other regarding an apartment in Rio de Janeiro* has to be valid. Only valid legal statements are individuals present in the model. There is no invalid *legal statement*. This follows the foundations of jurisprudence discussed in this section. We will denote as *contract* the legal statement (1). Let us denote by BR the set of (valid) individual

legal statements in Brazil, and, by SC the corresponding set regarding to $Scotland$. Since 18 is the legal age in Brazil, there is no individual legal statement about Peter in Brazil. On the other hand, the statement *Maria is of legal age*,$Maria-l-age$ for short, is in BR, and $Peter-l-age$ is in SC. There is a natural precedence relation between legal statements, only legally capable individuals have civil obligations. In other words, $contract \preceq Peter-l-age$ and $contract \preceq Maria - l - age$. Let PIL_{BR} be the set of legal statements in Brazil describing its *Private International Law*. Of course we have $PIL_{BR} \sqsubseteq BR$. By its very nature, PIL_{BR} is a disjunction of sets of legal statements subsumed by $\exists LexDomicilium.ABROAD$. It is worth noting that *Private International Law (PIL)* relates legal statements in different contexts, locations, time, etc. Thus each member of PIL regards a specific context, here we deal with geographical living place. $ABROAD$ is the union of the legal statements holding in each country, but Brazil. *LexDomicilium* is a legal connection, a relationship between laws in jurisprudence terminology. The pair of legal statements $\langle Peter - l - age, Peter - l - age\rangle$ is in *LexDomicilium*, since Peter lives in Scotland, abroad Brazil. Summing up, we have:

$$
\begin{array}{l}
Maria - l - age \in BR \\
Peter - l - age \in SC \\
contract \preceq Peter - l - age \\
contract \preceq Maria - l - age \\
PIL_{BR} \sqsubseteq BR \\
SC \sqsubseteq ABROAD \\
\exists LexDomicilium.SC \sqsubseteq \exists LexDomicilium.ABROAD \\
\exists LexDomicilium.ABROAD \sqsubseteq PIL_{BR} \\
\langle Peter - l - age, Peter - l - age\rangle \in LexDomicilium
\end{array}
$$

Thus, from what was discussed above, we can conclude that $contract \in BR$, for each legal statement generalizing $contract$, with regard to $\preceq$, namely $Peter - l - age$ and $Maria - l - age$, is in BR. For the interesting case we note that $Peter - l - age \in \exists LexDomicilium.SC \sqsubseteq PIL_{BR} \sqsubseteq BR$, by the definition of $\exists R.C$ concepts.

If one uses ALC instead of iALC in the above example formalization, she/he will need to consider a legal ontology involving non-valid *Legal Statements*, and hence an *ad hoc* ontology regarding jurisprudence main concepts. Dealing with non-valid legal statements will increase a lot the complexity of

the ontology considered also. Of course we simplified our example, since it only considers $Peter - l - age$ and $Maria - l - age$ as succeeding $contract$. In a real ontology, many more statements would have to be considered, $Maria - owns - the - apartment$ is among them. This simplification would turn much more complex the ALC case than the iALC.

Concerning the scalability of our approach, it can be argued that it scales as well as, or better than the traditional one, based on Classical ALC. Our approach does not have to deal with concepts outside jurisprudence, or more related to the application of laws. Dealing only with valid legal statements is a way to avoid describing the auxiliary terms and their ontologies. On the other hand, our approach forces us to relate these valid individual legal statements according their intrinsic juridical aspect. This can be done by considering subsumption between "sets" of individual laws. Stating $A \sqsubseteq B$ entails that each individual law a of kind A is preceded by every individual law b of kind b. By using our approach one has to consider only kinds of individual laws and their precedence relationship. Anyone building a legal ontology, using traditional way, has to perform this task also. Moreover, by using the traditional way of building legal ontologies, one has to deal with terms and concepts outside jurisprudence. This is, basically, the (potential) additional effort that traditional ontology practitioners have to deal with.

5. Conclusions

In this article, describing work-in-progress, we used iALC, a constructive description logic, to provide an alternative, and more adequate, definition for subsumption, that copes with the jurisprudence theory that views "The Law" as all (possible) legally valid individuals laws, instead of the totality of all (existing) valid laws.

An example of conflict of laws, namely geographic conflict of laws, was formalized by means of iALC in order to show its adequacy to perform coherence analysis in legal AI. We compared our approach with the more traditional approach, based on classic ALC, to the ontological formalization of "the law" that includes the development of one or more domain ontologies to be used in validating the legal statements. A brief discussion on the scalability comparing both approaches was done.

Finally, we have to says some words about the state-of-the-art in performing reasoning in iALC . In (dePHR) it is presented a Sequent Calculus that provides a basis for the design of an automated reasoner for iALC . This Sequent Calculus is based on the labeled Sequent Calculus for ALC presented in (RHP2009). This is the basis for the development of a tools for performing reasoning on legal ontologies built under our approach.

References

Bentham, J. *An introduction to the Principles of Morals and Legislation*. Athlone Press, London, 1970.

Bozzato, L., M. Ferrari, P. Villa *A note on constructive semantics for description logics.* Accepted at CILC09 - 24-esimo Convegno Italiano di Logica Computazionale.

Braüner, T and de Paiva, V. *Intuitionistic hybrid logic.* J. Applied Logic (JAPLL) 4(3):231-255 (2006)

Kelsen, Hans. *General Theory of Norms*. Clarendon Press, Oxford, 1991.

Hart, H. *The Concept of Law*. Clarendon Press, Oxford, 1961.

Hofmann, M. *Proof-theoretic Approach to Description-Logic.* In Proc. of Logic in Computer Science (LICS-05), 229–237, 2005.

Hohfeld, W. *Fundamental Legal Conceptions as Applied in Legal Reasoning*, Yale Univ. Press, 1919. Fourth printing, 1966.

de Paiva, V. *Constructive description logics: what, why and how.* Technical report, Xerox Parc, 2003.

de Paiva, V., Haeusler, E. H., Rademaker, A. *Constructive Description Logic: Hybrid Style*. Hybrid Logics 2010.

Michael Mendler, Stephan Scheele. *Towards Constructive DL for Abstraction and Refinement.* Description Logics 2008.

Rademaker, A., Haeusler, E. H. , Pereira, L. C. . *On the Proof Theory of ALC.* In: Walter Carnielli; Marcelo Coniglio; Itala D'Ottaviano. (Org.). The Many Sides of Logic. London: College Publications, 2009, v. 21, p. 273-285.

Raz, Joseph. Legal Principles and the Limits of Law. *The Yale Law Journal*, 81:823–854,1972.

Simpson, A.*The Proof Theory and Semantics of Intuitionistic Modal Logic.* PhD Thesis, University of Edinburgh, December 1993, revised September 1994.

Valente, A. *Legal knowledge engineering: A modeling approach*. IOS Press, Amsterdam, 1995.

An Ontological Representation of EU Consular Law

Erich Schweighofer
University of Vienna, Centre for Legal Informatics

Abstract. At present, EU consular law is under legal scrutiny by the European Commission. The CARE study reveals good pragmatic application but also significant implementation problems. As a site effect of our analysis, we have developed a concept of a legal ontology for knowledge description, multilingual information retrieval and semi-automatic application of consular law using a dialogue system. First experiments show the potential of this approach.
Keywords: Consular assistance, diplomatic protection, EU law, legal ontologies, dynamic electronic legal commentary, dialogue systems, multilingual information retrieval

1. Challenges of EU Consular Law

Article 23 of the Treaty on the Functioning of the European Union (TFEU) gives every citizen of the Union the right to consular and diplomatic protection if his or her Member State is not represented in a specific third country. Whichever mission (of another EU member state) the EU citizen ends up asking for support, the mission has to provide support on the same conditions as for their own nationals.
Article 46 of the Charta on Human Rights lays down the same right. The Green Paper "Diplomatic and consular protection of Union citizens in third countries", presented by the Commission in 2006, focuses on strengthening this right: In it, the European Commission points out that European citizens are not fully aware of this right, and that the legal consequences of it are far from being fully implemented by the Member States. After the consultation phase of the Green Paper, the Action Plan 2007-2009 "Effective consular protection in third countries: the contribution of the European Union" was adopted. One important measure is the examination of Member States' legislations and practices on consular protection and the assessment of the extent and nature of the observed discrepancies between Member States.
The CARE (Citizens Consular Assistance Regulation in Europe) project (http://www.careproject.eu) aims at offering tools to the Commission which support the European Commission in performing this examination. The CARE database collects relevant legal materials on diplomatic and consular protection adopted in each EU Member State. Various types of documents are collected: legislation, case law, administrative directives and guidelines, and also other informative materials made available by national governments for their citizens. The database contains full text documents in their original language, enriched by a metadata set, i.e. information about

the documents. Metadata are translated into English and French. Texts of the most relevant documents are translated into English and French as well. The database is accessible by all European citizens via the Internet (http://www.careproject.eu/database). A comprehensive report analyzes the legal framework in the EU Member States based on assessments of 27 national correspondents.

From a legal point of view, significant insufficiencies of implementation of Article 23 TFEU exist, in particular concerning legal frame work, standards of legal rules, reimbursement etc. These problems are solved in practice with a pragmatic implementation.

An ontological analysis shows that conceptualisation of consular law remains sketchy. Neither International treaties nor national laws have developed a strong terminology on consular law. Even a lexical ontology may provide important assistance.

Further, an ontology can be considered as an approach for solving the problem of multilingual (e.g. in 23 Community languages) handling of consular cases (see for the long list of functions Article 5 of the Vienna Convention on Consular Relations), taking into account the 27 different consular protection laws and policies. The ontology can provide required equivalence of concepts but can be linked also to a dialogue system.

For these reasons, experimental research on legal ontologies and dialogue systems has been undertaken. The remainder of this paper is organized as follows: Section 2 describes the consular law legal information system, section 3 the ontology of EU consular law, section 4 the dynamic legal electronic commentary, section 5 first experiments and, last but not least, in section 6, tentative conclusions are presented.

2. Legal Information System CONSUL

Handbooks in paper have long ceased to constitute best practice for dissemination of information. Websites and information systems are able to very nicely present the complex knowledge while coping very efficiently with often daily updates (e.g. travel recommendations). For finding materials, legal search constitutes an indispensable tool. Legal retrieval remains the best solution for determining the similarity between documents and queries (Manning et al 2008, Turtle 1995, Schweighofer 1999). For smaller domains like consular law with a complex structure, hypertext systems are a powerful tool. The flexible way of access with a non-linear representation of knowledge allows a user-friendly access to this body of knowledge.

The existing CARE database already allows full text information retrieval and browsing in the document collection. Our more powerful document retrieval system is going to be built using Apache Lucene (Apache Lucene

2010, Gospodnetic & Hatcher 2006), which offers state-of-the-art text retrieval capabilities but also allows fine-tuning of the information retrieval system according to the document properties of our text collection. Apache Lucene fulfils also the requirement of easy maintenance of the text corpus but also an efficient handling of the various versions.

3. Ontology of EU Consular Law

Since the 1990ies, ontologies as a conceptualisation of a domain are considered as tool for organising legal knowledge. Later, the idea of a semantic web (Berners-Lee 2001) with a mark-up that makes the text intelligent and active energized the concept of legal ontologies. For a long time, the University of Amsterdam has set the standards of legal ontologies with LRI-Core and now LKIF (Hoekstra, Breuker, De Bello/Boer 2007). Legal ontologies were implemented for tasks of conceptual information retrieval, knowledge representation, multilingual information retrieval or exchange of information and knowledge (see (Casanovas et al. 2007) and (Casellas et al. 2009)).

In our case, we consider using two ontologies: a lexical ontology like in the LOIS project (Dini et al. 2005) and a much more developed Dynamic Electronic Legal Commentary Ontology (Schweighofer 2006, Schweighofer 2010a) (see below).

A thesaurus for indexing contains a list of every important term in a given domain of knowledge and a set of related terms for each of these terms. A lexical ontology builds up from this basis with works on glossaries and dictionaries, extends the relations and makes this knowledge computer-usable in order to allow intelligent applications. Lexical ontologies provide this formalized description of a domain that can be understood and re-used by a knowledge system.

Based on already existing indices and sketchy conceptual structures, a lexical ontology CONSUL with about 200 legal and factual descriptors with definitions and relations has been established. Content relations will be taken from standard WordNet relations (especially hyperonymy and hyponymy). For all concepts, an ILI will be created in order to support multilingual use but also multilingual retrieval. Methodology is mostly derived from the previous LOIS project.

4. Dynamic Electronic Legal Commentary (DynELCom) CONSUL

The Dynamic Legal Electronic Commentary (DynELCom) (Schweighofer 2006, Schweighofer 2010a) CONSUL consists of a textual, e.g. syntactic representation of consular law that is supplemented by a semantic representation of the legal rules (e.g. conceptual representation of rules), a

semantic representation of the world (e.g. conceptual representation of facts) and a legal link structure between these repositories of knowledge. Knowledge acquisition is supported by semi-automatic text summarisation and text classification. A sketchy inference machine allows automated reasoning in "easy cases". A dialogue system establishes the facts but also handles the interface with the citizen.

The easier formalisation of knowledge and semi-automatic knowledge acquisition allow dynamic semi-automatic updating of the knowledge base.

The goal is an ontological index like that in legal commentaries, however, without the textual components. It is obvious that the readability of such ontological structures is limited and will require some training. However, the exact representation of the underlying conceptual und logical structure of the legal system is much better represented.

The DynELCom CONSUL is a model of a semantic legal knowledge system. Legal knowledge is formalised with tools of the semantic web and of legal ontologies. Browsing and handling of the legal text corpus is supported by a conceptual structure with links.

The main difference to existing approaches of legal ontologies lies in the fact that world ontologies (e.g. consular factual situations) are also included in this conceptual structure. As many quite developed ontological representations of world knowledge already exist, such knowledge can be used for enrichment of an ontological representation of the legal system.

A major part of the DynELCom CONSUL consists of the link structure between the facts (world ontology) and rules (legal ontology). Thus, legal reasoning is supported that may be sufficient in "easy cases and a valuable support in contradictory situations.

The formalisation of a legal knowledge domain with the DynELCom CONSUL allows also semi-automatic and automated applications. Conceptual search of links to factual and legal concepts are obvious results of this representation. This search can be supported by dialog systems that support the user in establishing relevant facts of a case. Thus, a sketchy form of automated legal reasoning can be offered, e.g. a "simplified legal syllogism". The facts of a case are properly refined by a dialog system leading to a factual concept but also a legal concept.

The DynELCom CONSUL faces the dynamics of the legal system. The indispensable indexing and analysis process is supported by semi-automatic categorisation and text analysis. Computational linguistics, text extraction, document categorization and text summarization tools are now sufficiently powerful so that good results can be achieved in very short time.

The analysis of the DynElCom is based on a co-operative work model between the man and the computer. The legal information system provides the basis for the commentary. The knowledge base with the ontology and semi-automatic text analysis provides extensive knowledge of the text

corpus of the legal information system. Software tools are information retrieval, hypertext, knowledge management, text summarisation, text categorisation and the inference machine. Manually, ontologies have to be established and maintained, semi-automatic indexing must be constantly fine-tuned and inference engines must be supervised. Such work is presently done by legal authors and practitioners. With the DynELCom CONSUL, a concentration of such analysis takes place in a semi-automatic way. The main advantage is real time delivery, higher quality and lower costs.

The main advantage of the DynELCom CONSUL seems to be obvious: in "easy cases", much of the work can be automated. Consular services would become cheaper with higher quality. Existing pressures on public budgets may lead to cuts in consular networks. With semi-automatic systems, much work can be outsourced to other consular posts of other Member States or honorary consuls.

Text corpus: The basis for the text corpus is the CARE project database. Only few modifications are envisaged; mostly hypertext links to the ontology, visual representations and a list of document types.

Ontology CONSUL: The ontology consists of a legal ontology, a world ontology and links (anchors) between the legal and the world ontology. Elements of the ontology are 3 types of frames: legal frame, fact frame and anchor frame. A frame contains a header, definitions (with sources), classification codes, and relations (to other frames, e.g. synonym, homonym, polysem, hyponym, hyperonym, antonym etc. but also to an anchor). The anchor frame can best be described as a citation with a header, the identification (abbreviation or number) and links to facts and legal concepts. For the representation, existing standards of the semantic web and legal ontologies are implemented; in particular OWL, RDF and LKIF.

This first step with a frame-like representation of legal concepts, factual concepts and the anchors between facts and rules will be followed by a second step that intends a more sophisticated ontological representation of the legal system. This representation focuses on space, persons, actions, material rules and procedural rules.

Action space for persons will be the real space and the cyberspace. Persons can be natural persons (and quasi-persons, e.g. robots or software agents). Objects in the space are physical objects (thinks), energy and quasi-physical objects (e.g. web store). Actions can be physical processes (actions or non-actions in real space or quasi-physical processes (actions on the web). Mental objects and mental processes consist of combinations of these elements. Due to social practice, such "virtual" sets are considered as a unified object or process (e.g. organizations, enterprises, associations, families etc.) Law builds on the existing physical and social structure of persons, objects and processes but modifies it or adds particular elements.

Persons can be natural or legal (e.g. limited company, international organization, state), objects are physical, mental or legal, and actions are physical, mental, or legal. It is obvious that the differences between the social reality and law (as representation as the world should be) are a high interest in any legal system. The representation is structured in concepts (an ontology), rules and factual situations. Isomorphism is respected via direct links to norms but also its logical representation. Legally relevant links between the world ontology and the legal ontology provide support for legal reasoning (e.g. possible factual situations or legal consequences of certain facts). In the LKIF terminology, such a function is called anchors (LRI-Core Ontology). They provide an anchor function as links between the social spectrum of actions and legally.

Knowledge acquisition tools: Text extraction and summarisation tools are decisive for the knowledge acquisition. The tools consist of a knowledge base containing the extraction, summarization and classification rules with header, rule, definition and relations and several tools for semi-automatic text analysis providing information on relevant documents, extract important text passages, classify documents, deliver definitions etc.

We have developed prototypes and applications on corpora-based text analysis for about 20 years now. Due to space restrictions, we can provide only a very short overview of the methods. A pre-defined list of descriptors can be checked against a text corpus with the KONTERM method (Schweighofer 1999). The various term occurrences are clustered according to the context allowing a structuring of homonyms and polysems. Thus, the various meanings in the text corpus can be analyzed. The self-organising map is a general unsupervised tool for ordering high-dimensional data in such a way that alike input items (e.g. documents) are mapped close to each other (Schweighofer 1999). In such a map, similar documents are grouped together. An extension allows the building of various layers and clusters of the map (growing hierarchical self-organising map). Further, common similarities of a cluster can be described with keywords (labelling of self-organising maps). Further, we have also taken advantage of the GATE library for text analysis. The GATE ANNIE (A Nearly New Information Extraction System) tool is very helpful for a more detailed analysis: segmentation of documents (tokenizer), words, gazetteer, sentence splitter and semantic tagger. The GATE JAPE tool (Regular Expressions Over Annotations) is implemented for a similar purpose (Gate 2010).

Sketchy inference engine: In first step with a simplified ontology, the inference is not much more than a hint of relevance like in the information retrieval system. Factual concepts are matched with legal concepts and vice versa. In case of a more complex ontology, an inference engine is required. Decision trees are represented as complex IF-THEN-statements with a mechanism for prioritizing rules. Such statements are interpretations of

facts, rules, concepts and anchors in the ontology. Such an inference engine allows the representations of a legal syllogism and a quicker handling of relevant information.

Dialogue system: Such a system is intended to converse with a human in a coherent structure (Wikipedia: Dialogue Systems 2010, Schweighofer 2010b). In the beginning, the dialogue will be text-based with a graphical user interface. A spoken dialogue system is in consideration. Natural language understanding is supported by a robust parser. The purpose of a dialogue system consists in the establishing of facts but also in the clarification of applicable legal rules.

5. Establishing the Ontologies and First Experiments

Due to time and financial restrictions, the implementation of the DynELCom CONSUL has mostly remained a concept. However, due to our ongoing involvement in the CARE project, we have worked for more about one year on a partial experimental application. The following presentation provides first experiments.

Existing text corpora (RIS, EUR-Lex, CARE) and ontologies resulting from our daily work with European, international and Austrian law forms the basis for these experiments. For Austrian and European law, we have established an ontology with a sufficient granularity of an ontological representation of a jurisdiction: about 10,000 thesaurus entries, 5,000 citations, up to 200 document types, a classification structure, 100 text extraction and summarization rules. This meta data is stored and updated in a database with different types of knowledge frames:

Fact and legal descriptors: header, definition (with sources), examples (with sources), relations (synonym, homonym, polysem, hyponym, hyperonym, antonym etc.), classification, other information.

Anchors: header, identification (abbreviation or number), synonyms, classification, author, links, other information.

Document types: header, identification (abbreviation), use, format, other information.

Classification: header, code, definition, relations, other information.

Extraction and summarization rules: header, rule, definition, relations, other information.

Concepts: header, definition (with sources), related thesaurus entries and citations, relations (synonym, homonym, polysem, hyponym, hyperonym, antonym etc.), classification, legal conceptual structure (ontological model), other information.

Rules: header, quasi-logical expression, source, type, classification, legal conceptual structure (ontological model), other information.

Procedures: header, flowchart, source, type, classification, legal conceptual structure (ontological model), other information.
This ontology was extended to the consular and diplomatic protection.
The following examples may show the lexical ontology (note: an (L), (F) or (A) is added to the header in order to distinguish between legal and fact descriptors as well as anchors). The attribute "legal conceptual structure" indicates relevant branches of law.

Legal concept:
```
Header: Evacuation (L)
Definition: In case of a catastrophe (e.g.
earthquake), EU Member States will evacuate their
citizens (and family members) as a matter of law
or policy. EU Member States will co-operate and
support each other for this goal (Art. 23 TFEU).
Source: Article 23 TFEU, national laws, CARE
project report
Relations: BT catastrophes (F), BT consular
assistance (L), catastrophes (A)
Classification: CAT:EVA
Legal conceptual structure: consular assistance,
catastrophes
Other information: none
```

Fact concept:
```
Header: 2010 Earthquake in Haiti (F)
Definition: Earthquake of 12 January 2010 with an
epicentre near the town of Léogâne affecting about
3 million people in Haiti.
Relations: Catastrophe (F), evacuation (L)
Source: English Wikipedia
Classification: CAT.EAR
Legal conceptual structure: Evacuation (L)
Other information: none
```

Anchor (link):
```
Header: Catastrophes (A)
Links: Terrorism (F), earthquake (F), tsunami (F),
hurricane (F), flooding (F), international
conflict (F), consular assistance (L), Article 23
TFEU (L), evacuation (L)
etc.
```

Figure 1: Examples of frames

At present, we are in the process of finishing the first prototype of this representation of Austrian consular law. The next step would be a verification of the conceptual structure using the knowledge acquisition and text analysis tools. Such a process is very time-consuming and requires financial resources not available so far. However, the existing ontology provides already a very helpful tool for legal work as it represents legal and fact concepts and its links.

6. Conclusions and Further Work

In this paper, we have given an outline of a system for semi-automatic application of consular law in a multilingual and multinational environment, focusing on the underlying legal ontology. For the moment, we are working on a more sophisticated and extended ontological representation.

Acknowledgements

The CARE Project is funded by the European Commission, Grant No. JLS/2007/FRC-1/50-30-CE-0226854/00-31.

References

Apache Lucene (2010), *High-performance, full-featured text search engine library.* http://lucene.apache.org/java/docs/ (accessed 14 February 2010).

Berners-Lee, T. et al.(2001), *The Semantic Web*, Scientific American Vol. 284, No. 5, pp. 34-43.

Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) (2007), *Proceedings of LOAIT 07,* II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, http://www.ittig.cnr.it/loait/LOAIT07-Proceedings.pdf.

Casellas, N., Francesconi, E., Hoekstra, R., Montemagni, S. (Eds.) (2009), *Proceedings of LOAIT 2009,* 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques joint with 2nd Workshop on Semantic Processing of Legal Text. Barcelona: IOT Series.

Dini, L. et al. (2005), *Cross-lingual Legal Information Retrieval Using a WordNet Architecture*, in: Proc. Int. Conf on Artificial Intelligence & Law 2005, ACM Press, New York, NY, pp. 163-167.

GATE (2010), *General architecture for text engineering,* http://gate.ac.uk/ (last accessed 20 June 2010).

Gospodnetic, O., Hatcher, E. (2005), *Lucene in Action,* Manning Publications.

Hoekstra, R., Breuker, J., De Bello, M., Boer, A. (2007), *The LKIF Core Ontology of Basic Legal Concepts*, in: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) "Proceedings of LOAIT 07, II. Workshop on Legal

Ontologies and Artificial Intelligence Techniques", pp. 43-64. http://www.ittig.cnr.it/loait/LOAIT07-Proceedings.pdf (accessed 16 May 2010).

Manning, C.D., Raghavan, P. & Schütze, H. (Eds.) (2008), *Introduction to Information Retrieval*. Cambridge, Cambridge University Press (2008).

Schweighofer, E. (1999), *Legal Knowledge Representation, Automatic Text Analysis in Public International and European Law.* The Hague, Kluwer Law International.

Schweighofer, E. (2010a), *Indexing as an ontological support for legal reasoning,* in: Yearwood, J. and Stranieri, A. (eds.), "Technologies for Supporting Reasoning Communities and Collaborative Decision Making", IGI Global (In print).

Schweighofer, E., (2006), *Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries,* in: Magnusson Sjöberg, D. and Wahlgren, P. (eds.), "Festskrift till Peter Seipel". Norsteds Juridik AB, Stockholm, pp. 569-588.

Schweighofer, E., Geist A. (2010b), *Semi-automatic application of consular protection law,* in: Schweighofer, E. et al., "Globale Sicherheit und proaktiver Staat – Die Rolle der Rechtsinformatik. Tagungsband des 13. Internationalen Rechtsinformatik-Symposions IRIS 2010". Wien, books@ocg.at (Band 266), pp. 509-513.

Turtle, H. (1995), *Text Retrieval in the Legal World,* Artificial Intelligence and Law, Vol. 3, Nos. 1-2, pp. 5-54.

Wikipedia (2010), *Dialogue systems*. http://en.wikipedia.org/wiki/Dialogue_systems (accessed 20 June 2010).

What do you mean? Arguing for Meaning

Tom M. van Engers[*], Adam Wyner[°]
[*]*Leibniz Center for Law, University of Amsterdam*
[°] *University of Leeds*

Abstract Building ontologies has been proven to be a complex issue in part because a community must commit to the conceptualization that the ontology represents. The community members must align their concepts and co-create. Arguing about a useful conceptualization is therefore an essential part of the process of designing an ontology. Logicians have developed formal argumentation theories, but have not combined formal argumentation with conceptualization. Rather, while conceptualization should play an important role in any argumentation theoretical approach, argumentation theories focus on arguments based on propositional logic and argument structures, which are not sufficient for arguing about domain conceptualization, which requires a more fine-grained logical analysis. In this paper we will explain why conceptualization plays an important role within argumentation and why argumentation support tools, especially if they use Natural Language Processing (NLP), can help in creating domain ontologies.

Keywords: Argumentation, Ontologies, Knowledge Acquisition, Natural Language Processing.

1. Introduction

Building ontologies has proven to be a complex issue in part because a community must commit to the conceptualization that the ontology represents. The community members must align their concepts and co-create. Arguing about a useful conceptualization is therefore an essential part of the process of designing an ontology. The creation of ontologies is usually done in small teams as part of informal knowledge engineering activities where participants discuss the conceptualization. Except where a minority has discretionary power to define the concepts, such a format is not suited for creating shared meaning between members of a larger community. However, in practice, people can cope with the task. For instance, where someone misunderstands, clarifying questions are asked and explanations given. Thus, the shared conceptualisation emerges from discussion; arguing about a useful conceptualization is an intrinsic part of communication. While it is not always easy for human beings to acknowledge and adjust to a different conceptualization, the problems of detecting conceptual differences and creating reconceptualizations are problems which are hard to solve in AI.

While one might expect that logicians working at formal theories on argumentation would have addressed the problems of conceptualization, thus far little attention has been paid to combining formal argumentation with conceptualization. Instead, argumentation theories focus on arguments

based on propositional logic, which is not fine-grained enough to argue about domain conceptualization.

Computational linguists have made significant progress in building ontologies from sentences expressed in natural languages. In order to address the hard AI problem of understanding natural language, researchers in this domain usually work with controlled languages (CLS). A good example of this approach is the Attempto Controlled English (ACE, see http://attempto.ifi.uzh.ch/site/description/), which is used by a relatively large number of computational linguists. Sentences expressed in ACE, i.e. in a somewhat restricted subset of the English language, can be parsed into first order logic (FOL) from which the ontology is derived.

One of the reasons to consider the interaction between natural language, argumentation, and conceptualisation is that knowledge engineers must translate from knowledge of a domain, often expressed in natural language, into a representation that is argued about. However, representing each sentence as a proposition hides crucial information that would help to relate statements or the contents of statements, draw inferences, filter redundancy, and identify contradictions.

In this paper we will illustrate why conceptualization plays an important role within argumentation and why argumentation support tools especially if they use Natural Language Processing (NLP) can help in creating domain ontologies.

2. Using CNL for Policy-making Discussions

We work with a scenario in which we want to support stakeholders to participate in policy-making discussions, using forum technology. For this purpose the domain knowledge, i.e. knowledge about the issues being discussed, must be made explicit, formal, and expressed in a language that a machine can process. This machine-readable knowledge representation we call the target form. Translating the knowledge that people have of a domain, which is often implicit, informal, and expressed in natural language, the source form, into the target form is a labour, time, and knowledge intensive task (see also Van Engers 2005), creating a "knowledge acquisition bottleneck" which has limited the adoption and use of powerful AI technologies (see Forsythe and Buchanan 1993).

In Wyner et al. (2010) we propose and outline a framework which extends multi-threaded discussion forums, integrating NLP, ontologies, and argumentation. The proposed framework goes beyond existing debate and argumentation support systems, by making the semantic content of the stakeholders in the policy-making debate formal and explicit. In this paper we will address the formalization rather than the construction of dialectical arguments.

While there are tools which support multi-user ontology development (see WebProtege http://webprotege.stanford.edu/) and there are ontology development tools which use natural language for input (see the AceWiki plug in for Protege http://attempto.ifi.uzh.ch/acewiki/), there is no support for arguing in natural language about an ontology. Rather, current ontology online multi-user systems such as WebProtege rely on the users to converge on an ontology or note the differences. Our proposal motivates the development of systems which not only captures the differences, but represents them as distinct ontologies for reasoning.

Broadly speaking, among the issues that need to be addressed are the following. Even if users enter in well-formed natural language sentences, how can we be assured that they enter in well-formed, meaningful rules for the formulation of arguments? Where we rely on input from public participants, who are not logicians or knowledge engineers with training in building well-formed rules, ill-formed arguments could be entered. This raises a general issue of what prompts can be introduced to make KB construction systematic and meaningful? For instance, at the level of propositions there is nothing incoherent about a rule such as *If P and Q, then R*. However, we see the rule is incoherent where P is *Bill is happy* and Q is *The Great Wall of China is long* and R is Swallows fly south in spring. Indeed, there is nothing preventing users from entering ungrammatical sentences, or sentences that are out of topic of the context of discussion. In the following we develop these issues.

One of the results and in some cases even one of the purposes of argumentation is to clarify issues by finding a shared conceptualization between the participants. Boer (in Boer 2009) citing Schlag (see Schlag 1996) stresses the importance of posing questions in (legal) arguments. He uses the following rhetorical hierarchy guiding those questions:

1. Ontological questions question the truth of terminological axioms and the ontological inferences based upon them.
2. Epistemic questions question the non-terminological inferences made from certain premises to certain thesis.
3. Normative questions address whether something is allowed or disallowed, good or bad etc.
4. Technical questions question the propositions of a case and are about the truth of the facts of a case.

This strength of attacking arguments depends on the rhetorical level, level 4 being the weakest and level 1 being the strongest attack.

In the following section we will explain some conceptualization issues that are relevant to argumentation.

3. Conceptualization issues in arguments

Participants involved in an argumentation process use natural language as the most import means of expressing themselves. In order to understand those expressions, the terms and syntactical information glueing them together has to be transferred into a conceptual model. Where the participants gradually come to understand one another, we have a process of shared conceptualization (Van Engers 2001). The shared conceptual model (ontology) only partly overlaps with the internal mental models of the stakeholders, and making an explicit conceptualization is usually a labour intensive task which requires lots of discussion because the (intended) meaning of concepts depend on the role those concepts play in the cognitive system of the individuals. Shared meaning has to be construed, requiring a 'rewiring' in the minds of these individuals. Mapping terms to a shared conceptualization can result in two typical inferential problems. The first one is class-referential mismatch, and the second is instance-referential mismatch.

An example of a class-referential mismatch is given in the following example where we have the following arguments:

Argument 1 consists of three statements in natural language;

Statement 1. People need a healthy living environment.
Statement 2. Plants are responsible for considerable air pollution.
Statement 3. Therefore plants should be prohibited in living environments.

Argument 2 also consists of three statements in natural language:

Statement 1. People need a healthy living environment.
Statement 2. Plants are responsible for regeneration of air.
Statement 3. Therefore we should have as many plants as possible in living environments.

Obviously the interpretation of these arguments would be quite different depending on what the concept would be that we want the term 'plant' to refer to.

An example of a instance-referential mismatch is the following. Suppose we have the following two arguments:

Argument 1:

Sentence 1: John is rich therefore John is happy.

and a rebuttal

Argument 2:

Sentence 2: John has severe health problems therefore John is not happy.

These arguments can be represented in the following AIF-graph:

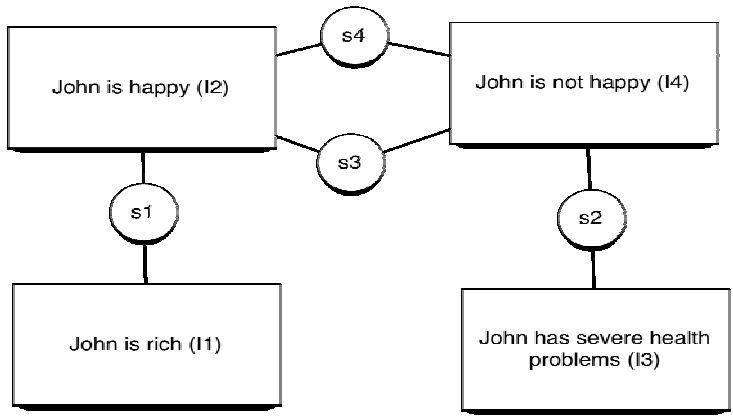

Figure 1. An AIF graph representing two conflicting arguments with a potential instance-referential mismatch. In this AIF-graph we'll find four I-nodes corresponding to

1. John is rich
2. John is happy
3. John has severe health problems
4. John is not happy

Obviously we expect that the John in all of these sentences refers to the same instance (assuming that this is what most readers will infer). But suppose that this is not the case and John in the first two I-nodes is referring to a different instance. In that case the two S-nodes representing the conflict between the second and fourth wouldn't make sense. In order to connect the I-nodes to the conceptualization we could use a mapping function. This mapping function would map the I-nodes 1 and 2 in our example to instance 'John12' and I-nodes 3 and 4 to John'34'. More precisely we would have two situations -- a situation before it was clarified that there are two Johns instead of one and the situation after this was clarified.

In the first AIF-graph the nodes would be functionally mapped to the same instance (John'12'). While in the second AIF-graph the I-nodes 1 and 2 in would be mapped to instance 'John12' and I-nodes 3 and 4 to John'34' and the S-nodes representing the conflict would be 'undercut' with a functional mapping to the 'exclusion' relation between John12 and

John34 in the conceptual model represented by the two sentences in our example.

Another conceptualization mismatch is the caused by the properties that individuals believe to belong to a concept. This problem could be solved to either split the concept in two or more concepts. This can be illustrated by the following example where we reuse the first argument of our previous example,

Argument 1 consists of three statements in natural language:
Statement 1. People need a healthy living environment.
Statement 2. Plants are responsible for considerable air pollution.
Statement 3. Therefore plants should be prohibited in living environments.

Argument 2 also consists of three statements in natural language:
Statement 1. Only some plants cause considerable air pollution.
Statement 2. Plants in living environments can help to reduce t travelling distance to work.
Statement 3. Therefore non-polluting plants should be allowed in living environments.

The second argument introduces a new concept (explicit in Statement 3) that of the non-polluting plant, which will require the splitting of the original concept plant into two concepts, one polluting plants, and another that of non-polluting plants. The reader must have detected the implicit argument in Statement 2 of the second argument that hides the conceptual relationship between travelling to work and air-pollution. Making this relationship explicit would require prompting in order to reveal all deductive steps implicitly made by the individual that made the statement.

The expressivity of AIF-graphs is intentionally limited to represent argument structures and not the content of the constituents of the 'I-nodes'. But this is unfortunately also the case in most other argumentation formalisation formalisms. Understanding the meaning of the arguments however does require a mechanism that allows for connecting the I-nodes to the corresponding conceptualization of the content of these I-nodes.

4. Conclusions and future work

In the IMPACT project we address argumentation in the context of policy modelling, which is a challenge. Firstly the participants in policy-making debates use natural language and understanding natural language is a hard AI problem. Secondly the dialectical form of the argumentation process may shift between different dialogue types (see e.g. Walton 1992).

Persuasion dialogue, information-seeking dialogue, negotiation dialogue, inquiry dialogue and sometimes even eristic dialogue can be mixed in such dialogues. We therefore have to limit the dialogue form and the language used, using a controlled language and a specific dialogue protocol in the forum.

On the argumentation formalization side we have little support yet either. The Dung framework (see also Laera et al. 2006) which we see as a basis for many argumentation theories is not typically useful in the context of policy making. In order to support the users in understanding the arguments, or policies, we need to be able to grasp the meaning of their expressions and give feedback about the consequences of their positions and choices. For this kind of feedback we have to go beyond the fourth level in the rhetorical hierarchy introduced in the section 3, i.e. the technical questions. We claim that in order to really support policy-making we need to be able to also cover the other rhetorical layers, up to understanding the meaning of the propositions, which implies that we have to formalise the participants' expressions using at least in FOL. We intend to further improve the NLP components as well as a component that can prompt participants posing rhetorical questions, as well as critical questions relevant to the argument (a plethora of papers on critical questions in argumentative settings can be found on Doug Walton's website see http://www.dougwalton.ca/papers.htm).

In our approach we hope to bridge between ontology building and argumentation theories which we believe is essential to both fields. As no knowledge will grow without arguments, we hope that our research contributes to more knowledgeable policy-makers and consequently to better policy.

Acknowledgements

The authors wish to thank the European Commission for sponsoring the IMPACT project.

References

Bailin, S. C., Truszkowski, W., 2002, Ontology Negotiation: How Agents Can Really Get to Know Each Other. In Proceedings of the WRAC 2002.

Boer, A., Van Engers, T.M., Van de Ven, S., 2010, Knowledge Acquisition from Sorces of Law in Public Administration, In: Proceedings of the EKAW2010 conference, under review. (2010).

Boer, A., 2009, *Legal Theory, Sources of Law, & the Semantic Web*. Dissertations in AI. IOS Press.

Van Engers, T.M., 2005, Legal engineering: A structural approach to improving legal quality. In Macintosh, A., Ellis, R., Allen, T., eds.: Applications and Innovations in Intelligent Systems XIII, Proceedings of AI-2005, Springer (2005) 3–10

Van Engers, T.M., 2001. *Knowledge Management: The Role of Mental Model in Business Systems Design*. PhD thesis, Vrije Universiteit Amsterdam, 2001. Belastingdienst.

Forsythe, D.E., Buchanan, B.G., 1993, Knowledge acquisition for expert systems: some pitfalls and suggestions. In: Readings in knowledge acquisition and learning: automating the construction and improvement of expert systems. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993) 117–124

Laera, L., Tamma, V., Euzenat, J., Bench-capon, T., Payne, T., 2006, Reaching agreement over ontology alignments, 2006, In *Proceedings of 5th International Semantic Web Conference* (ISWC 2006), Springer Verlag, p 371--384

Rahwan, I., Reed, C., 2009, The Argument Interchange Format, In: I. Rahwan, G. R. Simari (eds.), *Argumentation in Artificial Intelligence,* 383 DOI 10.1007/978-0-387-98197-0 19, c_ Springer Science+Business Media, LLC 2009

Russell, B., 1949, Human Knowledge – Its Scope and Limits, Allen & Unwin, London.

Schlag, P., 1996, Hiding the ball. *New York University Law Review*, 1681, 1996.

Walton, D., 1992, Types of Dialogue, Dialectical Shifts and Fallacies, In Van Eemeren, F., Grootendorst, R., Blair, J.A., Willard, C.A. (eds.), *Argumentation Illuminated,* Amsterdam, SICSAT, 133-147.

Wyner, A., Van Engers, T.M., Bahreini, K., 2010, From policy-making statements to first-order logic. In: Proceedings of eGOVIS 2010. (2010) to be published.

Wyner, A., Krasmir, A., Barzdins, G., Damljanovic, D., Davis, B., Fuch, N., Heofler, S., Jones, K., Kaljurand, K., Kuhn, T., Luts, M., Pool, J., Rosner, M., Schwitter, R., Sowa, J., 2010, On Controlled Natural Languages: Properties and Prospects}, In Fuchs, E. (ed.): Workshop on Controlled Natural Languages, CNL 2009, Springer, LNCS/LNAI 5972.

Ontologies, ICTs and Law
The International *Ontojuris* Project

Bibiana Luz Clara [1], Ana Haydée Di Iorio[2], Roberto Giordano Lerena[3]

1: Abogada. Profesora Investigadora de la Facultad de Ingeniería de la Universidad FASTA. Presidente del Instituto de Derecho Informático del Colegio de Abogados de Mar del Plata. Argentina. {bluzclara@ufasta.edu.ar}

2: Ingeniera en Informática. Profesora Investigadora de la Facultad de Ingeniería de la Universidad FASTA. Instructora Informática en Ministerio Público de la provincia de Buenos Aires. Argentina. {diana@ufasta.edu.ar}

3: Ingeniero en Sistemas. Profesor Investigador y Decano de la Facultad de Ingeniería de la Universidad FASTA. Argentina. {rogiord@ufasta.edu.ar}

Abstract. This article presents the experience of the International *Ontojuris* Project, modeled and developed to search and retrieve multilingual legal information based on ontologies and on the Universal Networking Language (UNL). It also presents the issue of multilingual information management, the importance of data processing from the semantic point of view and the possibility of semantic interoperability between systems, basically on Web search engines.

Key words: Ontology, Law, Artificial Intelligence, UNL, Ontojuris

Documentary Legal Informatics in Argentina

In the beginning, the development of legal informatics captured the attention of law practitioners to improve their working practices and increase accessibility to information and documents [1].

Due to the large amount of legal information in existence, it was necessary to find a support to facilitate access to this information, both to legal practitioners and citizens. The Documentary Legal Informatics would thus develop aiming at the automatic processing of legal information sources: legislation, jurisprudence and doctrine [2].

In Argentina, the best example of Documentary Legal Informatics is the Sistema Argentino de Informática Jurídica, SAIJ (Argentine System of Legal Informatics) created in 1979. The SAIJ is a government agency supervised by the Dirección Técnica de Formación e Información Jurídico-Legal (Technical Office for Legal Information), under the Subsecretaría de Justicia (Justice Subsecretariat) in the Ministerio de Justicia, Seguridad y Derechos Humanos (Ministry of Justice, Security and Human Rights). It provides normative, jurisprudential and doctrinaire information, whether national or provincial, taken from official sources.

The SAIJ[1] also coordinates the National Network of Legal Informatics, established in 1995. This net is constituted by all provinces that have signed agreements with the entity. Each province has a cooperation center in charge of providing and updating the provincial legal information[2].
In the Argentine legal system, jurisprudence is a formal source of law. For this reason, when a law practitioner carries out a jurisprudential search, he is seeking to reinforce the interpretation of standards or a personal point of view. In short, he attempts to present, by reference to verdicts, persuasive arguments to influence the judge's reasoning towards his side.
In addition to being limited by the syntactic search, most of these legal information search systems require the thorough knowledge of the verdict which the operator is trying to find: the year, actors involved, the court, and subject matter. At the moment of search, both in government initiatives and in private ones related to legal publishers, the legal practitioner frequently retrieves a large amount of irrelevant data that should be refined repeatedly until obtaining the desired result.
Many Artificial Intelligence (AI) techniques related to the representation of knowledge have tried to solve this problem. Among them, the representation by "ontologies" is noted; it refers to the formulation of a conceptual scheme within a given domain to allow the search of knowledge through meaning. This "ontological" representation is the basis for a real "Semantic Web" [3][4] by which the legal practitioner will be able to retrieve information from concepts, semantically, or by obtaining the exact data related to the search, all these independently from the possibility that in the referred text the specific term could be used at the moment of query [5][6][7].
In addition to the usual problems of legal search, the globality of law appears together with the complexity of varied conceptualization related not only to language but also to the particular cultures to which the concept refers.

Integration in a multilingual world

In a globalized, multicultural and multilingual world, access to resources is limited by multiple barriers, among them language and those originated in the interpretation of the real world to be conceptualized.

---

1 Source www.saij.jus.gov.ar – Institutional

2 There are also numerous initiatives of legislative, executive and judicial entities linked to the publication of legal information. Namely, the JUBA system of the Supreme Court of Justice of the Province of Buenos Aires, now accessible via the Web. It includes summaries and complete veredicts in the Province of Buenos Aires; the FANA system, also accessible on the Web including nationwide summaries and veredicts.

The use of ontologies, and conceptualization in itself, is complex, no matter which culture or language it is dealt with (or conceptualized). A higher level of complexity occurs when expanding the coverage of the proposed solution and the necessary conceptualization, to a multicultural and multilingual world.

Culture, terms, concepts, relationships between terms and term-concept relationships differ from one place to another, from one language to another. Whatever the domain, a given conceptualization that is valid in a certain language may not be acceptable in a different one. Thus, it is not possible to automate the process to a strict automatic translation, and, above all, it is impossible to homologate terms and concepts in different languages. There are words which cannot be translated in any language, simply because its strict meaning and use in its place of origin (in that specific culture) does not have a strict equivalent in another culture. Each term is the representation of a concept in a given language, and it may turn out that this concept will not have its equivalent in another language; hence, a possible word to represent that concept in that second language will not exist. Even in the same language (Spanish, for example), the same word may be used to represent different concepts in different countries or regions, and the same concept may be represented by different words in different regions using the same language.

The cultural complexity of languages and multilingualism are then transformed in a barrier to communication. In an interconnected and globalized world, it is urgent and necessary to undertake these issues from a technological point of view in order to facilitate intercultural communication.

Information on the web is growing daily and there is an urgent need to find "intelligent" searchers capable to work with semantics in the language and place where the search is performed. Users need to search by concept, not by term. Users think according to concepts, but must search the web for terms. The search is usually syntactic, not semantic. Browsers retrieve and return web pages containing search terms as they are spelt, textually, and not semantically. Some of them even propose pages in different languages where the specific term appears (having a different meaning in that language), without the ability to discriminate or prioritize on behalf of the concept.

There is also specific terminology in each domain or specialty which makes certain terms (words or set of words) have different meanings in a language, being the same country and culture. There are also concepts built on the basis of words that separately, have a certain meaning, but with a composition that does not mean the composition of those meanings. Traditional translators do not recognize this kind of compositions, namely, the representation of new concepts.

Users around the world need to see web pages from other countries, but on the basis of a specific concept, not terms representing it in their languages. This requires the development of search engines capable to understand and process the concept associated to the indicated term; with that concept (or meaning), engines should find pages having any of those terms or expressions representing it in their respective languages. These are known as intelligent search engines. They index and retrieve on-line meanings or concepts instead of words or terms. They include a conceptual infrastructure and ontological relationships that allow such management of search.

Given the importance of the term (actually, the concept) of the query, its vital correct interpretation, and since the above mentioned must be limited to the scope of law, the problem is increased. Misconception of a legal document is a very high risk that law practitioners cannot take; consequently, they need the support of technological tools to collaborate with their work and guarantee the correct interpretation of data, terms, information and documents involved in their decisions.

The UNL Program

In 1996, the United Nations General Assembly crated the Universal Networking Language Program (UNLP) as a project of the United Nations University (UNU).

The aim and activities of the UNLP are to develop and promote platforms and communication and information tools that will provide every nation the same opportunities to access, share and exchange scientific, cultural, social, and economic resources available in the global village. The UNLP has a flexible and dynamic net of persons and institutions devoted to developing, expanding, improving and multiplying the UNL System www.fi.unl.upm.es as a means of overcoming linguistic barriers; it is also a platform to collect and multiply human knowledge among people speaking different languages[3].

Ultimately, the project aims at allowing any person to share and retrieve information in their own language, no matter the language originating it. The project counted with an initial participation of 15 languages: German, Arabic, Chinese, Spanish, French, Hindi, Indonesian, English, Italian, Japanese, Latvian, Mongolian, Portuguese, Russian and Thai. The UNL System basically consists of UNL servers, UNL editors, and UNL viewers. The UNL language consists in UNL relationships and their attributes, universal terms, and a knowledge database[4].

---

[3] Source: www.unl.fi.upm.es

[4] source: www.undl.org

The Centro de Lengua Española (Spanish Language Center) and their group are working under the assistance of the Universidad Politécnica de Madrid (UPM) and they represent the Spanish language, not only in Spain, but also in every country sharing this language. Consequently, with this pretended universal program, the UNL language is adopted as the platform for the development of a multilingual legal server based on ontologies to pursue the International *Ontojuris* Project.

The *Ontojuris* Project

The International *Ontojuris* Project aims at facilitating a multilingual access to information about legal documents in the areas of Intellectual Property Law, Consumer Rights and Informatics Law. A consortium was then formed by researchers from Argentina, Brazil and Spain. Argentina was represented by Universidad FASTA; Brazil, by Instituo I3G; and Spain, by Univesidad Politécninca de Madrid. Experts from Universidad de Chile are also collaborating with the project.

The overall objective of the program consists in the research and development of an intelligent multilingual system based on ontologies, for the retrieval of legal information, limited in a first stage to the domains of Intellectual Property Law, Consumer Rights and Informatics Law [8].

Broadly, the stages in this project are as follows:

a) Selection of texts related to legislation, jurisprudence and doctrine of the domains involved in order to generate ontologies associated to each domain.
b) Identification and definition of terms and patterns of relationship. Construction of specific ontologies[5].
c) English determination of Headword[6] for each term.
d) Conversion of each term to the UNL and construction of domain terms inexistent in dictionary[7].

---

[5] The Project is developed under its own ontology editor, provided by I3G which anticipates the definition of some relationships.

The definition of ontologies was based on the identification of terms and by linking them after the following relationships: Synonymy, Type of (category or class), and Part (fraction or component).

[6] The UW (Universal Words) constitute the vocabulary of the UNL. They are concept labels, syntactic and semantic units that combine to form the UNL expression. Each UW represents a concept. A UW is formed by a Headword and a list of restrictions. The Headword may be a word, a compound word or a phrase in English. The list of restrictions is associated to the Headword to disambiguate and add specifications.

[7] The retrieval of universal words to fulfil the Ontology was achieved by referring to the UNL dictionary available at the Centro de Lengua Española. Those not included in the data base were created.

e) Definition of measures for indexing ontologies[8].
f) Definition of parameters allowing flow of ontologies.
g) Development of procedures for the integration of the ontology editor with applications and tools for the web search.
h) Modelling of the tool interphase.
i) Integration of the UNL to the ontology editor.
j) Expansion of search through the Universal Word (UW)[9].
k) Specification of the result presentation.

In developing the tool, the methodology of Knowledge Engineering, based on the semi-formal ontology description, was used to support the process of ontology engineering. In this methodology, instances of representation do not include the description of objects, only their relationships within a given domain.

The editor was designed to support the task and experience of the Knowledge Engineers when constructing the multilingual ontologies. It is a complex structure which connects terms taking into account concepts in the knowledge of their specific application. This allows the editor to determine the context of the documents in the query: they are contextualized.

The basic components of the ontology editor are: classes (taxonomically organized) and relationships (representing the type of interaction between concepts in a given domain). The ontology representation dos not use axioms or instances.

Discussion. Aiming at the future.

Having concluded the *Ontojuris* prototype, it is now time to verify its utility "in the field", with law practitioners from different countries validating the system.

---

[8] The *Ontojuris* system uses the Ontology created by the editor to index and retrieve information in the specified legal documents (laws, decrees, doctrine). The terms established by the method of creation of Ontology are used in the indexing process. The terms considered as relevant in a given document are added to the list of terms. On the other hand a list of words is generated from a dictionary of the natural language of the document. Hence, each document is labeled wiith the indices of terms and the words it contains.

[9] This phase, not yet completed, suggests a new expansion method based on domain ontologies and UW in order to retrieve multilingual information. This is an ongoing study based on the possibility of relying on a domain dictionary of UW for each natural language of the original documents. Each document to be searched is converted into a term vector and a word vector; besides it is mapped with a UW vector by which it is also indexed. That is to say, during indexing, the system converts each term into its corresponding UW, and it converts the original term into each different language associated to that UW.

Future steps must also be discussed. On the one hand, the expansion of the project to other disciplines given that this type of browser may be adjusted to a domain in any field provided that experts accomplish an accurate selection of ontologies.
On the other hand, members of other languages will be invited to the consortium to expand multilingual competences to other languages.
In the field of system integration, and given the level of knowledge embodied in the ontology of Law, it is necessary to work for an on-line integration with Law systems in Latin America certifying the semantic inter-operability among the systems.
Finally, it must be noted that the participation of the consortium Universities in the future course UNESCO "TECLIN" – Linguistic Technologies for Children Education in Aboriginal Communities will allow methodologies and the developed technology to extrapolate to other fields and contribute to the fulfilment of the United Nations goals for the millennium.
There is already a pilot project in Argentina for the recovery of endangered languages (particularly, the *Quechua*) which enables a "dialogue" between modern languages, such as Spanish, and aboriginal languages (declared World Heritage Site) favoring conservation. The project is developed on the same methodological and technological basis carried out for interaction between different languages in the field of Law. With very encouraging preliminary results, the Centro de Investigación CIPCO (CIPCO Research Center), La Buhardilla Foundation, in Tucuman, Argentina, is working in this direction with the support of the *Ontojuris* Universities.

Conclusion

After overcoming issues like the availability of digital information, connectivity and technical interoperability, cultural diversity and linguistics appear to be the real problems to reach a global knowledge society. It is at this point where technology of information has a fundamental role and an enthralling challenge.
The *Ontojuris* project reveals the potential of technological tools available to add up to the socialization of knowledge in the great "Global Village".

Acknowledgements

The authors are grateful to the members of the *Ontojuris* project of Universidad FASTA and to the members of the *Ontojuris* Project in Spain and Brazil. Also, to Tania Bueno, Sonali Bedin, Hugo Hoeschl, Cesar Stradiotto and Jesús Cardeñosa.

We would like to thank Ana Inés Cosulich for her assessment in the English version of this paper.

References

[1] Peñaranda Quintero (2001), *Iuscibernética, interrelación entre el derecho y la informática,* Ed. Miguel García e hijo, Caracas, Venezuela.
[2] Luz Clara, Bibiana (2001), *Manual de Derecho Informático,* Ed. Nova Tesis, Rosario, Argentina.
[3] Castells, Pablo, *La web semántica*, disponible en http://arantxa.ii.uam.es/~castells/publications/castells-uclm03.pdf. (accedida 2 de Mayo de 2009)
[4] Castells, Pablo*, Búsqueda semántica basada en conocimiento*, disponible en http://nets.ii.uam.es/publications/castells-fds08.pdf (accedida 15 de Abril de 2009)
[5] Pompeu Casanovas (2005), *Ontologías jurídicas profesionales. Sobre conocer y representar el Derecho*, disponible en http://www.leibnizsociedad.org/secciones/mater/pon/textos/ontologias_pompeu.pdf. (accedida el 17 de Junio de 2009)
[6] Abian, Miguel Ángel (2005), *Ontologías, que son y para que sirven*, disponible en http://www.wshoy.sidar.org/index.php?2005/12/09/30-ontologias-que–son-y-para-que-sirven. (accedida el 4 de Octubre de 2009)
[7] Burners Lee, Tim (2000), *Conference on the Semantic Web* disponible en http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html. (accedida el 20 de Octubre de 2009)
[8] Proyecto Ontojuris: Disponible en www.i3g.org.br/ontojuris/sistema.html

Author Index