Towards a High-level Controlled Language for Legal Sources on the Semantic Web

François Lévy[1], Adeline Nazarenko[1], and Adam Wyner[2]

[1] LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS, France
(francois.levy|adeline.nazarenko)@lipn.univ-paris13.fr
[2] University of Aberdeen, United Kingdom
azwyner@abdn.ac.uk

Abstract. Accessing legal regulations and legislation on the web should be improved by semantic annotation of legal rules. However, producing this type of annotations is challenging because legal language is difficult to parse automatically and any annotation involves a part of interpretation. We argue that an extended controlled language must be defined to address that issue. We present aspects of a controlled language required for such use and outline an incremental approach that supports the annotation/interpretation work to transform complex, legal natural language into a machine-parsable form.

Keywords: natural language simplification, semantic annotation, legal rules, controlled languages, semantic web

1 Introduction

Legislation and regulations are used to guide the conduct of many layers of social interaction. They specify what actions are proscribed, permitted, and obliged; they define legal terms, e.g. "British Citizen"; and they provide rules (as well as exceptions) that must be observed at risk of suffering a penalty. Legal documents are widely available on the web in various forms and from every institution and organisation. In terms of the semantic web, these documents should be accessed through semantic annotations (metadata). However, such annotations are generally only available for the general topic of a document, yet many relevant, important, and interesting queries relate to richer, more fine-grained, and complex information, particularly bearing on the rules in the law. Representing rules in such a standard form could considerably improve semantic access to legal texts on the web, which could be indexed, searched, and compared on the basis of the rules they contain.

To take advantage of rule annotations, we must bridge the gap between natural language (NL) sources and a machine-readable representation of rules.

It is widely appreciated that legal language is highly variable, complex, often ambiguous, and difficult for current Natural Language Processing (NLP) tools to handle [8]. To overcome that complexity, we propose a novel intermediate, incremental approach to the analysis and translation of legal language into a machine-parsable form, framing and exemplifying a *pattern language* approach. It is based on a Controlled Natural Language at a high level (hCNL), which supports folding the orginal text into a structured representation and unfolding it back into the orginal text. The recurrent patterns of legal language are exploited to structure complex legal expressions into simpler, more manageable subexpressions that can be reused in Semantic Web applications.

To ground our discussion and provide a running example, we use a corpus that was previously reported in [11], which is a passage from the US Code of Federal Regulations, US Food and Drug Administration, Department of Health and Human Services regulation for blood banks on testing requirements for communicable disease agents in human blood, Title 21 part 610 section 40.

> To test for evidence of infection due to communicable disease agents designated in paragraph (a) of this section, you must use screening tests that the Food and Drug Administration (FDA) has approved for such use, in accordance with the manufacturer's instructions. You must perform one or more such tests as necessary to reduce adequately and appropriately the risk of transmission of communicable disease.

This paper is a preliminary and partial proposal for hCNL, laying out some of the key concepts and processes along with an example. In the remainder of the paper, we outline our proposals for hCNL in Section 2, propose operations which support the transformations from legal text into hCNL in Section 3, and exemplify the approach in Section 4. Section 5 relates our proposal to several prior approaches, and 6 discusses tools and issues.

2 Towards a Specification of hCNL

In this section, we outline the type of language that allows legal professionals to express their interpretation of the source regulation. In Section 4, we present a worked example with the text introduced in Section 1. A CNL adapted to legal language must have at least:

- *Deontic modalities*: Each donation **must** be labeled ...
- *Conditional rule statements*: **If** you ship autologous donations ..., (**then**) you must assure ...
- *Main verbs*: You are not required **to test** ...
- *Semantic roles, e.g. Agents and Themes*: **You** are not required to test **blood components** ....
- *Lists*: ... unless you meet **the following conditions**: **(A)** ...; **(B)** ...
- *Exceptions*: **Except for autologous donations**, you must label such human blood and blood components ...
- *Anaphora to various referential types, e.g. NPs, lists, and exceptions*: You must appropriately label such blood or blood components as required under paragraph **606.121** ...

Indexation primarily requires normalisation of vocabulary and statements along with isomorphism to the source text. However, a full translation of the text into an hCNL is out of reach due to the subjectivity of interpretation and the complexity of the source texts. Furthermore, we can allow different degrees of granularity of translation, e.g. one translation may be more fine-grained than another, though both serve as annotations of the same underlying text. On the other hand, normalisation ought to provide output statements that are parsable, support semantic analysis, and are useful for semantic web applications.

Thus, we have an hCNL that adheres to the following "ideology":

- The target language should be normalised and consensual.
- The target language should be rich enough so as to preserve the constituent structures of the legal text and be comprehensible.
- The input fragments of text ought to be relatively short.
- The target language should not stipulate an interpretation of the source text, which is determined by the analyst.
- For expressions with implicit information, the normalised form also ought to represent that implicit information, e.g. diathesis, obligations, etc. with implicit arguments.
- For complex sentences with subordinate clauses, derived structures must maintain the interpretation of the source.

In Section 4, we provide some sample hCNL expressions that realise this ideology.

3 Annotating legal content with hCNL

To facilitate the correspondence between the source text and expressions of hCNL, we suppose a tool, under a human interpreter's control, to support the process of systematic rewrites which transform the source text into an annotation statement in hCNL. The rewriting process is a cascade of analyses, where lower level constituents are identified then reused in higher level constituents or for anaphora resolution. The tool links the original fragments of text with the resulting controlled statements, whatever the number of rewriting steps required.

Lexical simplification. To reduce the complexity and ambiguity of long and complex noun phrases, we identify domain specific terminology. Complex terms are interactively added to the hCNL vocabulary and replaced in the source text.

Shallow syntactic analysis. POS-tagging and shallow parsing for NPs, VPs, and PPs identify main grammatical components. A rule-based approach extends the analysis to rule statements, exceptions, and lists.

Sentence simplification and normalization. Using indicators that signal relationships amongst clauses, e.g. subordination and list structures, source sentences are analysed into component constituent statements. Particular attention is given to issues of semantic scope, e.g. negation and modal operators.

Anaphora resolution. Anaphoric expressions are linked to candidate antecedents, letting the user select the appropriate one. Unresolved expressions indicate the need for further lexical and/or syntactic simplification.

The output statements ought to be parsable and semantically representable by a CNL tool similar to ACE, but extented to the hCNL constructs. Any

Definitions

BLOOD-OR-COMPONENT(s) $=_N$ [human] blood or blood component(s)
INFECTION $=_N$ evidence of infection
DISEASE $=_N$ communicable disease
DISEASE-AGENT(s) $=_N$ DISEASE agent(s)
LIST-OF-PAR-A $=_N$ list of DISEASE-AGENTS of paragraph (a)
S-TEST(s) $=_N$ screening test(s)
FDA $=_N$ Food and Drug Administration (FDA)
MAN-INSTRUCTIONS $=_N$ the manufacturer's instructions
AD-APPROPRIATELY $=_{ADV}$ adequately and appropriately

1st level of annotation

To test for INFECTION due to DISEASE-AGENTS contained in LIST-OF-PAR-A, you must use S-TESTS that the FDA has approved for such use, in accordance with MAN-INSTRUCTIONS. You must perform one or more such tests as necessary to reduce AD-APPROPRIATELY the risk of transmission of DISEASE.

Fig. 1. Term creation and substitution

remaining ill-formed segment would be reanalysed at lower levels of analysis. The output of these processes is a normalized set of expressions that satisfy the hCNL definition for rule statements. Each statement is associated to a unique identifier for referring purposes. The presence of undefined terms leaves room for interpretation but we can nevertheless query and compare the annotations.

4 A Worked Example

We apply the above guidelines and methodology (Sections 2 and 3) to the example of Section 1, transforming the source into an annotated and normalized form. We first define terms identified by Termostat [3] in the hCNL vocabulary, rewriting the source text as in Fig. 1. The resulting text is then parsed, which highlights the chunks and grammatical keywords, paving the way for further simplification as in Fig. 2. Anaphora resolution is often a complex process that requires to make explicit new terms. Antecedents for anaphoric expressions are identified, e.g.“you” referring to “you, an establishment that collects...” and “such use” referring to the testing event. The last step consists in normalizing the previous output to get a sequence of standardized statements as in Fig. 3. This final transformation is done manually to comply to the hCNL target language, although some pattern-based analysis might be applicable.

5 Related Work

There have long been efforts to translate from natural language into executable programs with a range of approaches and some successes [8]. In [7], legislation was manually translated in logic programs. CNLs have been applied to legal language [9]. *Semantics of Business Vocabulary and Business Rules* (SBVR) provides elements of a pattern language to reconstruct legal language in a systematic manner [6]. LegalRuleML is a mark-up language that is designed to

Highlighted text

[To test for INFECTION due to [DISEASE-AGENTs contained in LIST-OF-PAR-A]], [you must] use [S-TESTs that the FDA has approved for such use], [in accordance with MAN-INSTRUCTIONS]. [You must] [perform one or more such tests as necessary to reduce AD-APPROPRIATELY [the risk of transmission of DISEASE]].

Definitions

LISTED-DISEASE-AGENT(S) $=_N$ DISEASE-AGENTs contained in LIST-OF-PAR-A
LISTED-INFECTION $=_N$ INFECTION due to LISTED-DISEASE-AGENTs
RISK-OF-TRANSMISSION $=_N$ risk of transmission of DISEASE
TEST $=_V$ test for LISTED-INFECTIONs
APPROVED-S-TESTs $=_N$ S-TESTs that the FDA has approved for TEST
ESTABLISHMENT $=_N$ establisment that collects BLOOD-OR-COMPONENTs

2nd level of annotation

To TEST, ESTABLISHMENT must use APPROVED-S-TESTs, in accordance with MAN-INSTRUCTIONS. ESTABLISHMENT must perform one or more APPROVED-S-TESTs as necessary to reduce AD-APPROPRIATELY the RISK-OF-TRANSMISSION.

Fig. 2. Additional term substitution and anaphora resolution

Final and 3rd level of annotation

To TEST, it is obligatory for an ESTABLISHMENT to use APPROVED-S-TESTs.
If an establishment uses APPROVED-S-TESTS, it is obligatory that it uses the APPROVED-S-TESTS in accordance with MAN-INSTRUCTIONS.
It is obligatory for an ESTABLISHMENT to perform as many APPROVED-S-TESTs as necessary to reduce AD-APPROPRIATELY the RISK-OF-TRANSMISSION.

Fig. 3. Fragment normalization

represent legal rules for the semantic web [1]. [11, 10] translate directly from legal language into web-based semantic representations. However, those languages do not provide high level annotations for source texts.

Wide-coverage parsers with semantic representation have been applied to legal language [10], including long and complex sentences. *Oracle Policy Modelling* (OPM) system [2] is a suite of interconnected tools to parse sentences and make rule-bases available online. SemEx supports the annotation of business regulations by business rules through an iterative rewriting process [5, 4].

Each of these approaches have strengths and weaknesses. Often, chunks of text must be provided (as we do) or problematic constructions are scoped out of the solution. Parsers are of limited capability, and most CNLs are not of sufficient richness to directly represent legal language. Where semantic representations are available, they must be checked for correctness. Annotation of legal text has made progress, but needs deepening and extension.

6 Discussion

A fully automated translation to hCNL is out of reach. However, we can apply several tools to support progressive transformation. Standard ngram and term extraction tools can identify terminology at the first level. Shallow parsing highlights relevant constituents. Tools can be developed to indicate discourse and subordination structures. Machine learning and pattern mining might be applied to recognize common patterns used in a domain or type of text.

Given the limitations of current technologies, hCNL provides several advantages. An expert can use a tool to interactively and iteratively identify and develop key terminology. The constituent structure of sentences is preserved, while related to its constituents. Different levels of abstraction are available, from local phrases to propositions.

However, there are also issues and limitations of hCNL. Exception clauses, which can be complex, must be identified and associated with the rule. The specific meaning or function of terms must be maintained, e.g. "following". The size and structure of the lexicon must be constrained since there is a potential to proliferate distinctions. Care must be taken to modify the source *salva veritate*. Finally, the scope of quantifiers, modals, and negation must be carefully maintained to preserve both the interpretation and anaphoric links.

References

1. Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., Wyner, A.: OASIS LegalRuleML. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (ICAIL 2013). pp. 3–12. Rome, Italy (2013)
2. Dayal, S., Johnson, P.: A web-based revolution in Australian public administration. Journal of Information, Law, and Technology 1 (2000), online
3. Drouin, P.: Term extraction using non-technical corpora as a point of leverage. Terminology 9(1), 99–115 (January 2003)
4. Guissé, A., Lévy, F., Nazarenko, A.: From regulatory texts to BRMS: how to guide the acquisition of business rules? In: RuleML 2012. Montpellier, France (2012)
5. Lévy, F., Guissé, A., Nazarenko, A., Omrane, N., Szulman, S.: An Environment for the Joint Management of Written Policies and Business Rules. In: Grégoire, E. (ed.) ICTAI. vol. II, pp. 142–149. IEE-CPS, Arras, France (Oct 2010)
6. OMG: Semantics of business vocabulary and business rules (sbvr). formal specification, v1.0. Tech. rep., The Object Management Group (2008)
7. Sergot, M., Sadri, F., Kowalski, R., Kriwaczek, F., Hammond, P., Cory, T.: The British Nationality Act as a logic program. Communications of the ACM 29(5), 370–386 (1986)
8. Wyner, A.: Logic in the Theory and Practice of Lawmaking, chap. From the Language of Legislation to Executable Logic Programs, p. To appear. Springer (2015)
9. Wyner, A., Angelov, K., Barzdins, G., Damljanovic, D., Davis, B., Fuchs, N., Hoefler, S., Jones, K., Kaljurand, K., Kuhn, T., Luts, M., Pool, J., Rosner, M., Schwitter, R., Sowa, J.: On controlled natural languages: properties and prospects. In: Proceedings of the 2009 conference on Controlled natural language. pp. 281–289. CNL'09, Springer-Verlag, Berlin, Heidelberg (2010)
10. Wyner, A., Bos, J., Basile, V., Quaresma, P.: An empirical approach to the semantic representation of law. In: Proceedings of 25th International Conference on Legal Knowledge and Information Systems (JURIX 2012). pp. 177–180. IOS Press, Amsterdam, The Netherlands (2012)
11. Wyner, A., Peters, W.: On rule extraction from regulations. In: Atkinson, K. (ed.) Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference. pp. 113–122. IOS Press (2011)