An Empirical Approach to the Semantic Representation of Laws

Adam WYNER [a,1], Johan BOS [c], Valerio BASILE [c], and Paulo QUARESMA [b]

[a] *University of Liverpool, Department of Computer Science, United Kingdom*
[b] *University of Evora, Department of Computer Science, Portugal*
[c] *CLCG, University of Groningen, The Netherlands*

Abstract. To make legal texts machine processable, the texts may be represented as linked documents, semantically tagged text, or translated to formal representations that can be automatically reasoned with. The paper considers the latter, which is key to testing consistency of laws, drawing inferences, and providing explanations relative to input. To translate laws to a form that can be reasoned with by a computer, sentences must be parsed and formally represented. The paper presents the state-of-the-art in automatic translation of law to a machine readable formal representation, provides corpora, outlines some key problems, and proposes tasks to address the problems.

Keywords. Legislation, Parsing, Semantic Representation

Introduction

To automatically process laws that are expressed in natural language, they must be made machine-readable. There are several approaches to making legal texts machine-readable, depending on the goals and purposes to be served by the processed text. Among the approaches, legal texts may be processed to link documents, to annotate for information extraction, and to translate them to a formal representation. In this paper, we focus on the last approach, rendering text into a machine-processable formal representation, e.g. First-order Logic (FOL) formulas, that can directly be fed into existing automated deduction engines to check for consistency and redundancy as well as to draw inferences.

In this paper, we outline a work in progress, providing new preprocessed corpora, reporting novel observations on automated parsing and semantic representation of legal text, and identifying research targets for ongoing work to improve current parsers and semantic representations. Thus, the paper takes an *empirical approach* to linguistic phenomena and tool development. In the following, we first discuss some of the approaches and issues in Section 1, pointing out the limitations of current analyses and tools. In Section 2, we outline our method and tools along with the corpora. Observations are reported in Section 3. In Section 4, we discuss the research and outline future steps.

[1]Corresponding Author: University of Liverpool, Department of Computer Science, Ashton Building, Ashton Street, Liverpool, L69 3BX, UK; E-mail: adam@wyner.info

1. Background

One of the early ambitions of artificial intelligence and law was to develop expert systems and executable representations of legal knowledge. Several large scale projects were carried out e.g. [11,2], where relevant textual portions were manually identified, decomposed, paraphrased, then formalised in Prolog. In addition to the labour of analysis, these approaches had no methodology or linguistic framework. Since this early work, some commerical products have become available [5]; the tools can parse selected, preprocessed sentences with limited semantic expressivity. However, as the system can be served on the Internet, they are widely deployed for calculation of benefits or taxes.

There have been recent efforts to apply natural language processing tools to legal text, with limited success. One approach uses machine learning to classify documents or sentences rather than fully parse and semantically represent text, e.g. [7] for Italian. Other approaches use linguistically oriented parsing, though are of limited scope [10], are not successful with legal text [12], or are not available for English [6].

Previous efforts have not succeeded on a wide scale. Statistical models can deal with the combinatorial explosion problem caused by ambiguities in natural language; and fast and robust processing tools have been produced based on such models [3]. Yet, large, richly annotated, open-source *gold standard* corpora are needed on which to build the statistical models. While such gold standards exist for other classes of text, e.g. newswire articles, no such corpora exist for legal texts; consequently, statistical models sufficient to accurately process legal text cannot be built. Thus, to make further progress, it is essential to develop a gold standard corpora. As part of this endeavour, the capabilities of current processing tools must be closely evaluated and modified as relevant.

2. Method

In this section, we briefly describe the tool and materials that we use in our study. We use C&C/Boxer [3], a tool that parses and semantically represents text. We have used this in the context of the Groningen Meaning Bank (GMB) [1], which is a freely available, online corpus of English texts that have been automatically parsed and given a semantic representation using the C&C/Boxer tool. GMB comprises thousands of public domain documents. In addition, GMB supports an open approach to curating a gold standard corpus of syntactic and semantic representations, for the texts and their preprocessed analyses are available to users to comment upon and to revise.

C&C/Boxer consists of a combinatory categorial grammar (CCG) parser [3] and Boxer, a tool that provides semantic representations in Discourse Representation Structures (DRSs) of Discourse Representation Theory (DRT) [9] . A categorial grammar specifies typed categories of lexical items along with their mode of combination: for example, a verb such as *runs* has the category of NP/S, and a noun such as *Bill* has a category of NP; combining something of NP category on the left of something of NP$\backslash$S category on the right yields a category of type S, which is a sentence. Along with the syntactic parse, a formal semantic translation expressed in the λ-calculus is provided, where the semantic derivation follows the structure of the syntactic parse. DRT was developed to provide semantic representations for discourses, including pronominal anaphora and discourse relations. The semantic representations are, by and large, FOL expressions that are suitable for FOL theorem provers and model builders; some sentential operators (e.g. possibility and necessity) are also represented.

Syntactic and semantic derivations can be given for long, complex sentences and discourse continuations, and we discuss one sample below. However, sentences and discourses must be carefully checked that the derivation is correct and, more importantly, that the semantic output corresponds to semantic intuitions for an interpretation of the meanings of the sentences. Thus, we use the tools to analyse, curate, and then commit statements to a gold standard corpus, while also improving the tool.

Turning to discuss the materials, we have created a new, automatically parsed and semantically represented (using C&C/Boxer) corpus of legal text; the output must be carefully checked and modified in order to form the basis of a gold standard corpus. Currently, this corpus contains *British Nationality Act 1981, Part I (BNA1)*, where irrelevant or ideosyncratic aspects have been removed (e.g. HTML and layouts). Comparing a corpus of Voice of America texts (VOA) to BNA1, we find that: BNA1 has more tokens on average per sentence than VOA (77.9 vs. 21.4); and BNA1 has greater average complexity than VOA (4.39 vs. 3.09), measured as the average recursion level of a DRS.

3. Observations

In this section, we comment on the parse and semantic representation of one example:

> *If an application is made to register as a British citizen a person who is a British overseas territories citizen, the Secretary of State may, if he thinks fit, cause the person to be so registered.*

By and large, C&C/Boxer correctly parses and semantically represents this sentence: it identifies the conditional structures, scopes the modal "may" over the lowest consequent clause, identifies predicate-argument structures, and locates antecedents for pronouns (e.g. "he" is the Secretary of State) amongst other aspects.

However, there are issues, which we give along with proposed solutions.

- An abstract noun "application" is given as the agent of the action of "registering", rather than an individual. Abstract nouns should be analysed in terms of a verb.
- The phrase "he thinks fit" is misanalysed to represent that by thinking, one is fit. Verbs with sentential complements should be weighted in the parser.
- The agentive interpretation of "cause" is not semantically represented; that is, the Secretary of State is not given as the agent who causes the registration. Causal verbs ought to impose their thematic structure.
- The ellipsis introduced by "so" does not refer to "as a British citizen", as it should. Ellipsis could be resolved in terms of syntactic anaphora.

These problematic constructions do not seem to appear in newswire texts, which explains why a parser trained on them performs poorly on them. Having identified difficult linguistic features, we can develop strategies for better parsing and semantic representation.

4. Discussion and Conclusion

In this section, we point out a range of additional topics for future work. First, while C&C/Boxer successfully parses and semantic represents complex sentences with FOL, there are constructions that are understood to go beyond FOL, e.g. *generalised quantifiers*, *genericity*, and *intensional operators*, among others. Initially, we could provide a semantic representation for these, leaving the full logical and model-theoretic aspects underspecified. Second, there are well-known issues about references to sections or in-

dividuals across a legal text, e.g. *the following conditions* [11]. This requires an augmentation of the tools that resolve anaphora. Third, we have not discussed inference or contradiction. We would take the approach of Recognising Textual Entailment tasks (RTE) [4], using the *Nutcracker* inference engine provided by C&C/Boxer. For this, we must first develop a gold standard corpus of statements in legal language where entailment and contradiction relations hold between textual portions. Finally, it is widely acknowledged that defeasibility, where an inference may not hold in some particular circumstance, is essential to legal reasoning [8]. In our sample sentence, this appears in the phrase "if he thinks fit". Currently, C&C/Boxer relies on notions of FOL strict inference. A task for defeasible textual entailment may be a useful approach.

While there are all these significant issues, we can make useful, systematic, incremental progress on the analysis of the expressions. Broadly, the advantage of our empirical approach is that we have a transparent, systematic, and grounded means to curate the corpus and modify the tool. In this paper, we have provided an automatically parsed and semantically represented corpus of legal text; the corpus provides the opportunity to develop a rich, articulated gold standard, which can then be used to train and thereby improve statistical parsers. In addition, the analysis indicates ways in which the parser and semantic interpreter can be revised to improve performance.

Acknowledgments

The first author was supported by the FP7-ICT-2009-4 Programme, IMPACT Project, Grant Agreement Number 247228. The views expressed are those of the authors.

References

[1] Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 92–96, Avignon, France, 2012.

[2] Trevor Bench-Capon, Frans Coenen, and Paul Orton. Argument based explanation of the british nationality act as a logic program. *Information and Communications Technology Law*, 2:53–66, 1993.

[3] Johan Bos. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications, 2008.

[4] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4), 2009.

[5] Surend Dayal and Peter Johnson. A web-based revolution in Australian public administration. *Journal of Information, Law, and Technology*, 1, 2000. Online.

[6] Emile de Maat and Radboud Winkels. Suggesting model fragments for sentences in dutch laws. In *Proceedings of Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2010)*, pages 19–28, 2010.

[7] Enrico Francesconi. Legal rules learning based on a semantic model for legislation. In *Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT-2010)*, Malta, May 2010.

[8] Jaap Hage. Law and defeasibility. *Artificial Intelligence and Law*, 11:221243, 2003.

[9] Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language: Formal Logic and Discourse Representation Theory*. Springer, 1993.

[10] L. Thorne McCarty. Deep semantic interpretations of legal texts. In *ICAIL '07: Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 217–224, New York, NY, USA, 2007. ACM Press.

[11] Marek Sergot, Fariba Sadri, Robert Kowalski, Frank Kriwaczek, Peter Hammond, and Therese Cory. The British Nationality Act as a logic program. *Communications of the ACM*, 29(5):370–386, 1986.

[12] Adam Wyner and Wim Peters. On rule extraction from regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 113–122. IOS Press, 2011.