



Introduction for artificial intelligence and law: special issue “natural language processing for legal texts”

Livio Robaldo¹ · Serena Villata² · Adam Wyner^{3,4} · Matthias Grabmair⁵

Published online: 13 April 2019
© Springer Nature B.V. 2019

The law has always been an attractive domain for language and semantic technologies since it is essential for governance, and it pushes the state of the art in natural language processing (NLP) to its limits.

Recent research has highlighted the need to create a bridge between conceptual questions, such as the role of legal interpretation in mining and reasoning, as well as computational and engineering challenges, such as the handling of big legal data and the complexity of regulatory compliance. To facilitate progress towards integrating the efforts on these two objectives, the EU has recently funded several research projects, among which ‘MIREL: MIning and REasoning with Legal texts’—<http://www.mirelproject.eu>).

Researchers in artificial intelligence and law have long worked to bring information mining and reasoning together. More recently, practitioners must effectively use sophisticated natural language processing technology on large volumes of publicly accessible legal texts so as to benefit to society as a whole.

The development of such NLP methods and semantic technologies for automatically analysing, indexing, and enriching big data that is freely available on the web has created opportunities for building new approaches to improve the efficiency,

✉ Livio Robaldo
livio.robaldo@uni.lu

Serena Villata
villata@i3s.unice.fr

Adam Wyner
a.z.wyner@swansea.ac.uk

Matthias Grabmair
mgrabmai@andrew.cmu.edu

¹ University of Luxembourg, Luxembourg city, Luxembourg

² CNRS, Inria, I3S, Université Côte d’Azur, Sophia Antipolis, France

³ Swansea University, Singleton Park, Swansea SA2 8PP, Wales, UK

⁴ <https://www.swansea.ac.uk/staff/law/adamwyner/>

⁵ Carnegie Mellon University, Pittsburgh, USA

comprehensibility, and consistency of legal systems. For example, one approach is to associate syntactic elements, e.g., nouns, verbs, and clauses, to their semantics, e.g., the individuals, properties, relations in a given domain. The meanings could be everyday as well as specific to legal information and are of special interest to practitioners in private practice, government, public administration, education, and research.

On the one hand, the EU has delivered vast amounts of resources on EU law in many languages in recent years (such as, EuroParl, JRC, etc.). On the other hand, mature NLP and Semantic Web technologies can be used to: automate extraction of knowledge from legal documents; formalize legal data as ontologies; and represent an ontology as Linked Data in RDF. The results of applying the technologies to the law can be used to may support many tasks that would benefit from structured data and automatic legal reasoning such as better search and information retrieval, compliance checking and decision support, as well as a better presentation of the legal information to professional and non-professional stakeholders.

This special issue addresses a range of NLP approaches applied to legal texts to address classification, knowledge representation, argument mining, information extraction, information retrieval, ontology population, and multilingualism in legal documents.

With the above in mind, we have selected the following papers for the special issue:

CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service

by *Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni.*

The paper presents an experimental study where machine learning (ML) is employed to automatically detect such potentially unfair clauses in terms of service of on-line platforms. The ML technology has been implemented in the CLAUDETTE (automated CLAUse DETectEr) tool (<https://claudette.eui.eu>), which evaluate compliance of online consumer contracts and privacy policies with personal data protection law, specifically the General Data Protection Regulation (GDPR).

Building a corpus of legal argumentation in Japanese judgement documents

by *Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga*

The paper presents an annotation scheme describing the argument structure of judgement documents, together with a manually annotated corpus containing 89 documents (37,673 sentences; 2,528,604 characters) and the first two stages of an implemented algorithm for the automatic extraction of argument structure. The annotation scheme is based upon blueprint models of summaries of various granularities, and it is intended to support summarisation aimed at the legal professions, a central construct in Japanese law.

Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora

by *Ilias Chalkidis and Dimitrios Kampas*

Deep Learning is one of the main NLP technologies used nowadays. The application of Deep Neural Networks has increased significantly also in legal analytics. The paper presents a survey of the early adaptation of Deep Learning for legal analytics inferences focusing on three main fields: (1) text classification, (2) information extraction, and (3) information retrieval. An online-available pre-trained legal word embeddings has been developed, using the word2vec model over large corpora, comprised legislations from UK, EU, Canada, Australia, USA, and Japan among others.

Unsupervised and Supervised Text Similarity Systems for Automated Identification of National Implementing Measures of European Directives

by Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Guido Boella, Lorenzo Grosio, Marco Gerbaudo, and Francesco Costamagna

The paper presents both unsupervised semantic similarity systems and supervised machine learning models, in order to infer national implementations (NIMs) of European directives. Previous works have proposed and utilized unsupervised lexical and semantic similarity techniques based on vector space models, latent semantic analysis (LSA) and topic models. However, these techniques were evaluated on a small multilingual corpus of directives and NIMs. The paper develops new word and paragraph embedding-based semantic similarity models from a multilingual legal corpus of European directives and national legislation. The performance of these models is compared with the previously proposed methods on a multilingual test corpus of 43 directives and their corresponding NIMs.

Semi-Automatic Knowledge Population in a Legal Document Management System

by Guido Boella, Luigi Di Caro, and Valentina Leone

The paper presents some NLP applications that semi-automate essential time-consuming and lower-skill tasks such (1) classifying legal documents, (2) identifying cross-references and legislative amendments, (3) linking legal terms to the most relevant definitions, and (4) extracting key elements of legal provisions to facilitate clarity and advanced search options. The use of natural language processing tools to semi-automate such tasks makes the proposal a realistic commercial prospect as it helps keep costs down while allowing greater coverage. The procedure has been implemented in the Menslegis system (<https://www.augeos.it/EN/DynamicContents/Details?ContentId=MENSLEGIS>), distributed by Nomotika SRL, a spin-off of the University of Turin, in collaboration with Augeos SRL (<https://www.augeos.it>).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.