LREC 2018 Workshop

1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph

PROCEEDINGS

Edited by

Georg Rehm, Víctor Rodríguez-Doncel and Julián Moreno-Schneider

ISBN: 979-10-95546-18-4
EAN: 9791095546184

12 May 2018

Proceedings of the LREC 2018 Workshop
1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph

12 May 2018 – Miyazaki, Japan

Edited by Georg Rehm, Víctor Rodríguez-Doncel and Julián Moreno-Schneider

http://legalkg2018.lynx-project.eu

Acknowledgments: This work has received funding from the EU's Horizon 2020 research and innovation programme through the contract LYNX (grant agreement no.: 780602).

Organising Committee

- Georg Rehm, DFKI GmbH, Germany*
- Victor Rodriguez Doncel, Universidad Politécnica de Madrid, Spain*
- Julián Moreno-Schneider, DFKI GmbH, Germany*
- Elena Montiel, Universidad Politécnica de Madrid, Spain
- Tatiana Gornostaja, Tilde, Latvia
- Jorge Gracia, Universidad de Zaragoza, Spain
- Martin Kaltenböck, Semantic Web Company, Austria
- Ilan Kernerman, K Dictionaries, Israel
- Matteo Zanioli, Alpenite, Italy

* Main editors and chairs of the Organising Committee

Programme Committee

- Adam Funk, University of Sheffield, UK
- Adam Wyner, University of Aberdeen, UK
- Alvaro Rodrigo Yuste, Universidad Nacional de Educación a Distancia, ES
- Christian Sageder, Openlaws, Viena, AT
- Emilio Serrano, Universidad Politécnica de Madrid, ES
- Enrico Francesconi, National Research Council of Italy (ITTIG-CNR, IT) and Publications Office of the EU, LU
- John Hendrik Weitzmann, Wikimedia Deutschland e.V., DE
- John McCrae, National University of Ireland Galway, IR
- Jorge González Conejero, Autonomous University of Barcelona, ES
- Jorge Gracia, Universidad de Zaragoza, ES
- María Navas, Universidad Politécnica de Madrid, ES
- Mariano Rico, Universidad Politécnica de Madrid, ES
- Milan Dojchinovski, University of Leipzig, DE
- Pablo Calleja, Universidad Politécnica de Madrid, ES
- Paloma Martínez, UC3M, ES
- Paulo Quaresma, University of Evora, PT
- Peter Bourgonje, DFKI GmbH, DE
- Philipp Cimiano, University of Bielefeld, DE
- Tomaso Agnoloni, Institute of Legal Information Theory and Technologies, IT

Preface

The World Wide Web and also Internet-based applications as well as startup companies are diversifying with an astonishing pace. While, only a few years ago, general-purpose applications and companies with a rather broad scope have been ubiquitous, now more and more niches and highly specific domains are being explored, both by academic and industrial research and also by enterprises and entrepreneurs.

One of the specific domains that has been receiving a lot of attention recently are legal and regulatory information systems for cross-border commerce, not only under the umbrella of the European Union's Digital Single Market but also in other areas and regions. Of special interest in that regard is the recent hype around "regtech" (i. e., applying technologies to regulatory applications), the use of cognitive computing and language technologies to tackle the law (especially regarding the extraction of structured information from legal documents) and the adoption of semantic technologies for publishing legislative documents by the European institutions. These developments clearly indicate that a new generation of intelligent applications is appearing in the legal domain.

These applications rely extensively on text resources and Semantic Web technologies. There is also a strong demand for high-quality, well described language resources, which can be used in the legal domain. Vast amounts of cases, rulings, laws, regulations, political programs, parliamentary debates and public opinions have been released in the last few years. The recently started research and innovation project Lynx, funded by the European Union, suggests to aggregate all the available information from the legal domain into the Legal Knowledge Graph. However, within the research community we still lack a clear consensus or agreement of what the key characteristics, components and functionalities of the Legal Knowledge Graph should be. The language resources to be used to populate the Legal Knowledge Graph are also a topic of current debates.

This workshop is aimed at the first steps towards filling this crucial gap. In order to develop new products and services that are meant to assist lawyers and legal experts, new types of interoperable language resources and technologies are necessary, aimed specifically at constructing and making use of the Legal Knowledge Graph.

The organisers of the 1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph would like to thank all contributors for their valuable submissions. We would also like to thank all members of the Programme Committee for reviewing the papers.

G. Rehm, V. Rodríguez-Doncel, J. Moreno-Schneider May 2018

Programme

Opening Session

14:00 – 14:20 Welcome and Introduction

Session 1

14:20 – 14:40 Elena Montiel-Ponsoda and Víctor Rodríguez-Doncel:
Lynx: Building the Legal Knowledge Graph for
Smart Compliance Services in Multilingual Europe

14:40 – 15:00 Julián Moreno-Schneider and Georg Rehm:
Towards a Workflow Manager for Curation Technologies
in the Legal Domain

Session 2

15:00 – 15:20 Gerardo Sierra, Gemma Bel-Enguix, Guillermo López-Velarde,
Ricardo Saucedo and Lucía Rivera:
Event Extraction from Legal Documents in Spanish

15:20 – 15:40 Akshay Minocha and Navjyoti Singh:
Legal Document Similarity using Triples Extracted
from Unstructured Text

15:40 – 16:00 Aikaterini-Lida Kalouli, Leo Vrana, Vigile Marie Fabella,
Luna Bellani and Annette Hautli-Janisz:
CoUSBi: A Structured and Visualized Legal Corpus of US State Bills

16:00 – 16:30 **Coffee Break**

Session 3

16:30 – 16:50 Damir Cavar, Joshua Herring and Anthony Meyer:
Case Law Analysis using Deep NLP and Knowledge Graphs

16:50 – 17:10 Julián Moreno-Schneider and Georg Rehm:
Curation Technologies for the Construction and
Utilisation of Legal Knowledge Graphs

17:10 – 17:30 Milagro Teruel, Cristian Cardellino, Fernando Cardellino,
Laura Alonso Alemany and Serena Villata:
Legal text processing within the MIREL project

Closing Session

17:30 – 18:00 Discussion and conclusions

Table of Contents

Case Law Analysis using Deep NLP and Knowledge Graphs

Damir Cavar, Joshua Herring, Anthony Meyer
Indiana University, Semiring Inc.
Bloomington, IN
dcavar@iu.edu, joshua@semiring.com, anthony@semiring.com

Abstract

We present a system for mapping facts and knowledge in legal texts, in particular case law opinions and holdings to knowledge graphs, enabling advanced semantic search over the case law corpus, as well as matching of case descriptions onto case laws using graph similarity. The essential components for knowledge graph generations are deep linguistic NLP components. We discuss how the deep analyses provided by these components allow us to process not only the core semantic relations in the legal documents, but also to process advanced semantic and pragmatic properties, including implicatures and presuppositions.

Keywords: Case Law, Deep NLP, Knowledge Graph

1. Introduction

In this paper we discuss ongoing research and development activities in the domain of Deep Linguistic Natural Language Processing (NLP) technologies for the analysis of legal documents with a focus on court decisions, opinions, case law, and case documentation. Leveraging deep linguistic annotation technologies like semantic and pragmatic preprocessing to map cases, case law and opinions onto knowledge graphs enables us to bring the power of graph matching search and concept-based reasoning – involving automatically detected implicatures and presuppositions – to the world of legal AI (see e.g. Potts (2015)).

The goal of this project is to provide linguistically-informed, detailed analyses of case law based on professional-grade comparisons between individual cases. In pursuit of this goal, we model the content of individual case documents with knowledge graphs (KGs) built from semantic and pragmatic processing during the text-mining step. These KGs can later be used directly to compare individual cases, saying how they reinforce, contradict, or build on one another, eventually providing human-quality analysis of the current state of the body of law as well as automatic detection of trends that may have escaped human observers, all at a fraction of the cost in time and resources.

2. Previous Work

There are numerous commercial and free tools to search and process case law files. We will not go into details here with the existing commercial solutions. To our knowledge, none of these commercial solutions seems to provide a deep content analysis that is supported by fine grained linguistic and semantic technologies.

There are various documented and publicly available knowledge graph and ontology implementations for legal applications. Soria et al. (2007) describe an ontology of (Italian) *law paragraphs*, i.e., fundamental units of codified law. At the highest level, each paragraph belongs to one of three classes: *obligations*, *definitions*, or *modifications*. These are divided into subclasses. The class *obligations*, for instance, contains the subclasses *obligation*, *permission*, *prohibition*, and *penalty*. They also train a classifier to assign these classes to law paragraphs automatically. They report high precision and recall scores (96% and 92%, respectively).

3. Data

At present, our source material consists of the corpus provided by the Free Law Project (`https://free.law/`), which is an interface and mirror repository for the Public Access to Court Electronic Records (PACER) (`https://www.pacer.gov/`) service provided by the Administrative Office of United States Courts (`http://www.uscourts.gov/`) to facilitate public electronic access to federal court records. The bulk of the data we survey comes from the Free Law Project's CourtListener (`https://www.courtlistener.com/`) service and takes the form of compressed JSON files representing individual cases, organized by jurisdiction.

The JSON objects include meta-information and multiple content sections, but there is frequently no explicit separation of the opinion/holding from fact-finding and other components of the case. To detect the opinions in the case files, we use Machine Learning (ML) approaches, training automatic classifiers on a sub-corpus manually annotated by legal experts to separate the holding from residual contextual information about the facts of the case.[1]

4. Architecture

The primary step in processing is to extract core semantic relations, e.g. subject – verb – object, from clauses in the text, which we do by means of NLP components. We use the Natural Language Toolkit (NLTK) (Bird et al., 2009) components for basic segmentation and tokenization, followed by Part-of-Speech (PoS) tagging and WordNet-based[2] hypernym, hyponym, and synonym annotation of

[1] We use Scikit Learn (Buitinck et al., 2013; Pedregosa et al., 2011) and additionally various text classifiers based on Bayesian or Support Vector Machine approaches.

[2] For details on WordNet see Miller (1995) and Fellbaum (1998)

nominal elements. Providing the extended taxonomic relations for allows us to index textual content such that concept search and search over synonyms is made possible of the case law corpus. This information is also essential for mapping of concrete concepts to concepts in KRs, as explained below.

The Stanford CoreNLP (Manning et al., 2014) pipeline provides extended analytical components, including lemmatizer, a constituent parser, a dependency parser, and a coreference analyzer. The spaCy pipeline (an implementation of Honnibal and Johnson (2015)) is comparable to CoreNLP, except that it does not have constituent parsing and coreference analysis components.

All these components face performance issues and tend to fail on complex sentences or sentences that exceed a particular length. By way of example, the following sentence will receive some linguistic annotation of very limited use:

> *Their attack is anchored in a Fifth Circuit case, United States v. Whitfield, which involved two state judges who were convicted of accepting bribes from an attorney in exchange for favorable rulings in his cases.*[3]

It is not uncommon for the types of constructions found in formal documents to be misanalyzed, particularly clause level coordination, constructions with ellipsis or gapping, empty subject constructions, and many other constructions which require processing that goes beyond the level of combining textually represented words into composite meanings. Rimell et al. (2009) and Nivre et al. (2010) report that even the best parsers perform quite poorly where unbounded dependencies are concerned, i.e., dependencies such that there is no theoretical limit on the distance between head and dependent. Often, parses for such constructions are unsystematic and unpredictable, and we are consequently forced to augment them using the sub-optimal linguistic output across various NLP pipelines. For example, consider the sentence

> *They tasted the specimens to identify them.*

which contains a purpose clause, namely, *to identify them*. In CoreNLP's analysis for this sentence, the matrix verb *tasted* to *identify* via `advcl` (adverbial clause), the Stanford Dependency category that includes purpose clauses (De Marneffe et al., 2014). However, if *tasted* is replaced by *were tasting*, CoreNLP returns a parse in which *were tasting* is related to *specimens* via `dobj`, which is turn related to *identify* via the tag `acl` (adjectival clause). In other words, a simple change in verb tense can result in a fundamentally different analysis that overlooks a key semantic relation.

In addition to freely available NLP components and pipelines, we make use of in-house technology and infrastructure. Within the Free Linguistic Environment (FLE) project (Cavar et al., 2016) we developed multi-word morphological analyzers using a two-level transducer framework as made available in the Foma morphology compiler (Hulden, 2009). In this way, we are able to generate Finite State Transducers (FST) that recognize single- and multi-word named entities and jargon specific to the legal domain, such as *amicus curiae*. The recognized terminology is annotated using a "legal" tag, as well as semantic sub-type information, wherever applicable. In our case, the terminology is augmented with domain specific tags, for example indicating that an expression like "FMLA"[4] is typical in the labor law domain, while "exclusive rights" indicates the copyright law domain. An FST can read in a term like "FMLA" and output an analysis listing its tag(s) and sub-tag(s), much like a two-level morphological analysis in the manner of (Koskenniemi, 1983).

To be able to generate deeper linguistic analyses that cover linked constituent structure, functional relations, and morpho-syntactic and semantic properties, we work with a (Probabilistic) Lexical Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001; Cavar et al., 2016) based parsing system. Such a parsing system generates syntactic structures that encode scope relations between sentential (or clausal) elements, which is essential in, among other things, semantic processing of quantifiers, time reference, and negation. The FLE project is related to the Xerox Linguistic Environment (XLE) (Crouch et al., 2011) project, which is the most significant and complete implementation of the Lexical Functional Grammar (LFG) framework (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001) in a grammar engineering environment. It comes with the additional advantage of being well-documented, in, among other sources, a grammar engineering textbook (Butt et al., 1999), official technical documentation (Maxwell and Kaplan, 1996), and various online material (Crouch et al., 2011).

The analytical strength of LFG-based parsers can be exemplified using a simple example. The sentence *They offer several justifications for this position*[5] would receive a constituent structure (or c-structure) analysis as in figure 1.

The corresponding functional structure (or f-structure) is given in figure 2.[6]

The two representations are linked, such that every tree node has a link to an attribute-value-matrix (AVM) in the f-structure. The f-structure provides a rich set of morpho-syntactic and semantic features that are extremely useful for higher level analysis of semantic relations between arguments. The c-structure provides a phrase-structure analysis that represents scope relations between the sentential arguments and modifiers. It is in particular essential for the processing of scope of negative elements, operators, and quantifier.

Such rich representational output allows us to map the con-

[3] United States v. Martinez-Maldonado, United States Court of Appeals, First Circuit Nos. 12–12 89, 12–1290 see `http://media.ca1.uscourts.gov/pdf.opinions/12-1289P-01A.pdf`.

[4] Family Medical Leave Act, Public Law No. 103–3, 29 U.S.C. Sections 2601–2654 (1993)

[5] See footnote 3.

[6] The output for the c- and f-structure in figures 1 and 2 was generated using the XLE-Web interface `http://clarino.uib.no/iness/xle-web`. See for example Meurer et al. (2016).

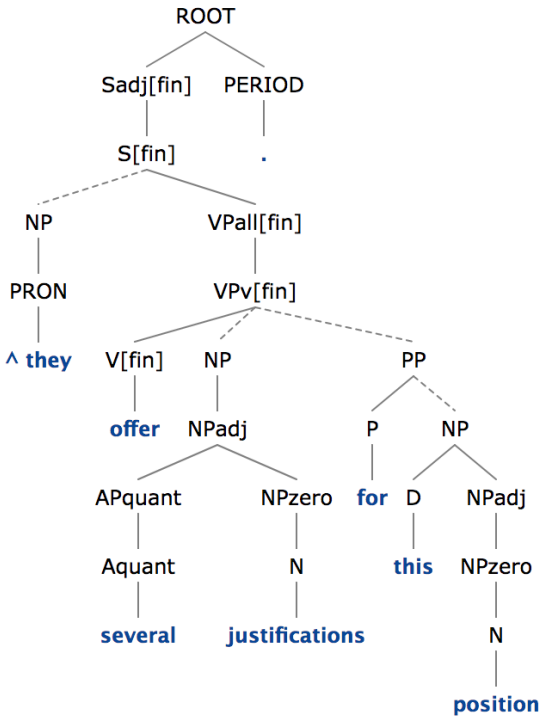


Figure 1: C-structure

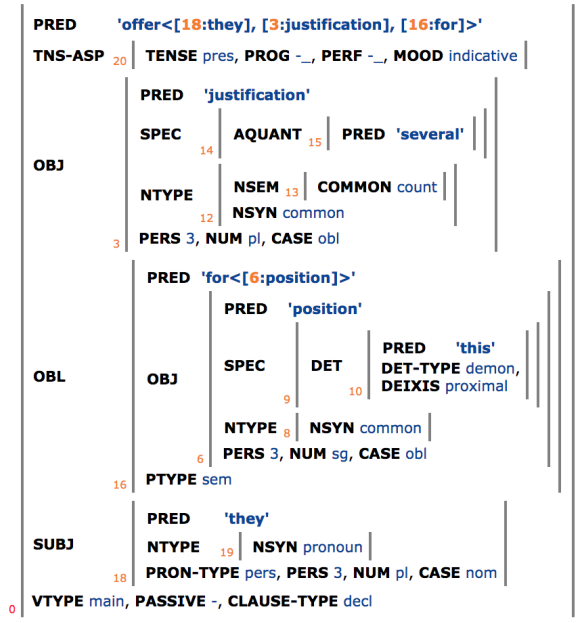


Figure 2: F-structure

tent of the target sentence to a KR by converting the main predicate to a triple, i.e. predicate *offer* – subject *they* – object *justification*. The subject *they* would be linked to a real referent (antecedent) in a given context using anaphora resolution and correference analyzers.

Additional properties for all arguments can be extracted from the f-structure in conjunction with sentential features, as well as dependency relations and functional roles of clausal elements.

To map content from unstructured text (a list of sentences and clauses), it is essential to identify the tense information, mood, or voice of the sentence, any scope relations between constituents and clauses within the sentence, and quantifiers or semantic operators, such as whether a negating element in a clause scopes over the entire clause or merely an isolated phrase, and whether it has scope over the matrix clause and thus potentially all embedded clauses in its scopal domain. It is also crucial to detect voice so as to determine the directionality of a relation in a tuple. The examples *The plaintiff accused the defendant of breach of contract* and *The defendant was accused of breach of contract by the plaintiff* would receive the same directionality representation of the KR predicate relation. This is obviously not true for a sentence like *The plaintiff was accused by the defendant of breach of contract*. Likewise, if the sentence uses future tense, it is not a factual or assertive statement that should be integrated into a common KR that represents factual knowledge, as for example in *Google will buy Apple*. Also, a past tense assertive statement embedded under a hypothetical or future tense matrix clause does not represent concepts and relations that should be part of a KR, as for example *We do not believe that Google bought Apple*. Our NLP components are capable of detecting mood, tense, and voice in the input sentences. Additionally, they can predict clausal structures of complex sentences as well as the scope relations between these clauses. This is in particular relevant, when it comes to the correct extraction and mapping of semantic relations in embedded contexts. If an embedded clause with assertive content of the kind *that the plaintiff transfered the funds to the bank account* can only be interpreted if the matrix clause as for example "*It is true*" is not negated, not using future tense or the subjunctive, and so on. The linguistic properties and the scope relations between the clauses are essential to be able to correctly differentiate hypotheticals from factive assertions, or guesses from wishful projections.

To extend this capability to legal language, we not only develop our own domain specific adaptations of NLP components and pipelines, we also post-process the outputs of the aforementioned openly accessible NLP components. The post-processing extends the linguistic analytical output and also corrects systematic errors of certain NLP components.[7]

In the FLE implementation we use a probabilistic model of the c-structure parser backbone as well as a probabilistic unification algorithm over Probabilistic Directed Acyclic Graphs (PDAG) for the AVM and f-structure representations. This specific version of an LFG-type of parser allows us to engineer more robust grammars that can cope with

[7] Some such errors are construction specific. Coordination of clauses as common in legal documents is systematically analyzed as local phrase coordination. Predicative modifiers as for example adverbial temporal constructions or prepositional location phrases in syntactic parse trees are frequently attached to adjacent noun phrases (low attachment), rather than the predicate. Many of these mistakes can be corrected using simple post-processing steps.

agreement violations or unification failures, as well as word order violations or complexities that other NLP pipelines cannot process. The grammars that underly such a parser can be engineered and trained using corpora and distributional models over syntactic trees and morpho-syntactic features.

Our architecture wraps all available NLP components in a Remote Procedure Call (RPC) (Microsystems, 1988) set of micro-services and lets each service process each input sentence in parallel. The outputs of each component are evaluated, scored and transformed into a synthesized unique Linguistic Data (LingData) data structure (or object). This LingData object decides on a uniform representation of the data, i.e. PoS-tags and dependency relations are normalized, clause boundaries are added or removed, etc. We implemented a specific extension that interprets constituent structure trees and dependency graphs into phrasal scope relations that allow us to query the hierarchical relations between all tokens, phrases and phrasal nodes in the sentence structure. This implies that we can not only identify a negation in a clause, but also the correct structural scope of it. In the sentence *the plaintiff did not violate corporate policies* the negation is correctly identified as sentential negation, while in the sentence *the plaintiff violated corporate policies and not federal law* the negation scope would be local over *federal law* only.

The resulting LingData objects thus contain complex linguistic properties and annotations. These are accessible using class specific methods. Currently the class is only available as a Python implementation. In future versions we will provide a C++ and a Go implementation as well.

The different LingData objects generated by parallel NLP pipelines are then combined into a single, hopefully complete and correct, analysis, using mapping of the linguistic analyses and extensions generated from the outputs of the NLP components in a Unification method.

Such a parallel architecture is complex and computationally expensive. It can, however, be easily scaled given our choice of a JSON-based RPC micro-services infrastructure regulated through a core processing dispatcher or manager. This infrastructure frees us from distracting programming language, operating system, or other technical dependencies, as components can be swapped, added, coupled, or removed as and when the need arises. For most of the advanced NLP components and pipelines the loading time of models is eliminated, since the components run in daemon mode and communicate over TCP/IP with the clients.

4.1. Knowledge Graph Mapping

As described above, the core semantic relations extracted by isolating the core predicate in a clause and its dependent functional phrases provide the core relational elements or sub-graph for the KR. Additional processing is necessary to map these attributes and properties to relations between concepts or concept attributes. For example, the construction *the plaintiff was employed as a clerk in the defendant's firm* could imply that the plaintiff is a clerk or that the concept of the asserted plaintiff in the KR has an attribute-value specification *profession – clerk*. We have an independent model for such mappings that allows us to generate graph relations for the specific domain or use-case.

The processing of semantic and pragmatic relations allows us to expand the KR even further. To be able to process implicatures or presuppositions, the NLP output needs to provide detailed information about nominal elements or phrases in the clauses. For example, if the direct object in a clause is a definite and specific noun phrase like *the plaintiff bought the blue car*, implicatures that can be generated to extend the KR representation of the situations and events would include factual statements like *there were multiple cars available that the plaintiff could have bought* and *no other of these cars is blue*. Likewise, a statement like *the plaintiff was petting his dog* presupposes that the statement *the plaintiff owns a dog* is true as well. While most of these implicatures and presuppositions will strike human readers as trivial, they can provide valuable information for automatic processing and KR generation. Moreover, some semantic and pragmatic side-effects so inferred might not be easily accessible to the reader at all – a situation that is particularly frequent in highly specialized, knowledge-based domains like legal reasoning.

For the processing of such relations we build construction-specific mappings for the domain, the particular language, and linguistic constructions in general. The mapping of definite and specific noun phrases to imply the existence of a super-set is an example of a linguistic property that can be applied universally in all linguistic domains. Other such mappings are language-specific and can depend on cultural peculiarities. By contrast, many of the semantic and pragmatic properties are domain-specific, and their specification and definition in specific NLP components requires the supervision and involvement of trained domain experts (legal professionals, for the present case).

At the graph level we use two commercial environments: Neo4J and Stardog. Both products are advanced graph databases with different capabilities when it comes to semantic processing. Neo4J serves as an experimental simple, but highly performant and scalable, graph representation system where we do not make use of extended semantic technologies like OWL-based ontologies (W3C OWL Working Group, 2012; W3C OWL Working Group, 2009) or reasoning (using a Description Logic framework).*See for example Antoniou and van Harmelen (2004).* Stardog, by contrast, functions as an extended graph with OWL-backing assertions of facts, concepts, relations, and attributes. We augment Stardog with Pellet (Sirin et al., 2007) as a reasoner. One goal is to use the ontology as a classification system for concepts that allows us to generate extended properties for asserted individuals. For example, if an ontology defines *CEOs* to be *humans*, and *humans* have *birthday*s, *gender* and *parents* as properties, when we assert that *John Smith is CEO*, the system can automatically extend the properties of the concept *John Smith* to include the implications (*John Smith*) *has a birthday*, *has gender*, *has parents*, etc. This level of semantic expansion using common reasoners (e.g. Pellet) augments potentially sparse assertions and makes hidden facts and circumstances explicit and available for search and graph-based comparison or analysis.

Another goal is to detect conflicts in assertions related to

the types of concepts. For example, if we assert that *Grungle Inc.* is a company, and the ontology encodes taxonomic relations or the concept hierarchy that a CEO is a (*isA*) human, an assertion of the type *Grungle Inc. is the CEO of Sprackets Inc.* can be flagged or rejected as a violation of base relations formulated in the ontological concept relations. While common OWL-based assertion handling would not be able to catch such violations, extensions of reasoners like Pellet can be used to detect conflicting assertions of this kind.[8]

Our analysis of different graph databases for the back-end storage of a knowledge graph in our system did also include the Apache Jena (`jena.apache.org`) environment. Due to obvious limitations here, we will extend the discussion of the suitability of knowledge graph storages for our purposes to subsequent publications. In addition to these free systems we made arrangements to evaluate other commercial graph database systems as for example TigerGraph (`www.tigergraph.com`), where our main interest lies in performance for search and graph comparison with large knowledge graphs.

5. Discussion

Given the limits of this article, it is of course impossible to provide an exhaustive list of the capabilities and advantages that deep NLP and semantic processing using Description Logic, OWL ontologies, and reasoning can bring to legal language processing. We hope we have nevertheless been successful in conveying the impression that they are prodigious, representing an evolutionary leap in applicative power. Mapping case law documents, in particular the opinion and the holding, to KRs allows us to search over the document base via graph similarity.

Mapping specific concepts to hypernyms introduces an conceptual abstraction layer that allows us to identify cases with concrete reference to for example *injury involving a semi-truck* can be found by searching for *injury involving a vehicle* or even *car*.

Mapping concrete case files onto a graph representation in a systematic way allows us to use graph similarity search to identify semantically related cases, holdings, and opinions. Using comparisons of graphs within a similar concept and relation space allows us to identify conflicting opinions in the case law, or conflicting facts in other document types. These types of conflict studies open up new possibilities for the automatic analysis of case law holdings and opinions.

We are aware of the fact that we owe the reader a detailed explanation of the architecture, performance, and issues related to the NLP components and architecture.

One serious problem for us in the current situation is that we do not have any objective measure for the performance of our system, due to the lack of gold standard resources and corpora. While we can describe the technical and runtime behavior, the accuracy of some NLP components, we cannot yet easily assess on a larger scale the extraction of semantic relations and concepts.

[8]The developers of Pellet and Stardog informed us that this is a possibility in their system, and we assume that this is missing in other non-OWL-based graph-databases. Such restrictions can also be implemented in the free and open Apache Jena environment.

Due to a lack of appropriate resources, our evaluation right now can only be based on a usefulness study with paralegals and law firm employees.

Due to space limitations, we defer the exposition and discussion of the results of experiments and concrete applications to the concrete conference presentation and subsequent publications.

6. Acknowledgments

We are grateful for useful comments, code and algorithm implementations, and suggestions related to the project to all our other team members, in particular Stefan Geissler, Matthew Rust, Jacob Heredos, Malgorzata E. Cavar, and Lwin Moe.

Many more colleagues have helped us with work on the FLE project. We are grateful to Ron Kaplan, Ken Beesley, Lionel Clément, Larry Moss, Mary Dalrymple, Agnieszka Patujek, Adam Przepiórkowski, Paul Meurer, Helge Dyvik, Annie Zaennen, Valeria de Paiva for many helpful suggestions, data sets, grammar samples, ideas, and comments. Numerous other colleagues have been supportive, providing suggestions and advise in various development phases. Our colleagues from the Kelley School of Business or the Maurer School of Law at Indiana University have provided valuable domain specific information and knowledge. Various local lawyers have helped us to understand the essential processes and tasks involved in legal research.

7. Bibliographical References

Antoniou, G. and van Harmelen, F., (2004). *Web Ontology Language: OWL*, pages 67–92. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Butt, M., King, T. H., Niño, M.-E., and Segond, F. (1999). *A Grammar Writer's Cookbook*. CSLI Publications.

Cavar, D., Moe, L., Hu, H., and Steimel, K. (2016). Preliminary results from the free linguistic environment project. In Doug Arnold, et al., editors, *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, pages 161–181. CSLI Publications.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE documentation. Online document at http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Number 42 in Syntax and Semantics. Academic Press, New York.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–92.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

Hulden, M. (2009). Foma: A finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Kaplan, R. M. and Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. Cognitive Theory and Mental Representation, pages 173–281. The MIT Press, Cambridge, MA.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form production and generation. *Publications of the Department of General Linguistics, University of Helsinki. Helsinki: University of Helsinki*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Maxwell, J. and Kaplan, R. M. (1996). An efficient parser for LFG. In *Proceedings of the First LFG Conference*. CSLI Publications.

Meurer, P., Rosén, V., and Smedt, K. D. (2016). Interactive visualizations in the iness treebanking infrastructure. In Annette Hautli-Janisz et al., editors, *Proceedings of the LREC'16 workshop VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 1–7, Portorož, Slovenia. ELRA.

Microsystems, S. (1988). RPC: Remote procedure call protocol specification. Request For Comment (RFC) 1057.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nivre, J., Rimell, L., McDonald, R., and Gomez-Rodriguez, C. (2010). Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 833–841. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Potts, C. (2015). Presupposition and implicature. In Shalom Lappin et al., editors, *The Handbook of Contemporary Semantic Theory*, pages 168–202. Wiley-Blackwell, 2 edition.

Rimell, L., Clark, S., and Steedman, M. (2009). Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 813–821. Association for Computational Linguistics.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semant.*, 5(2):51–53, June.

Soria, C., Bartolini, R., Lenci, A., Montemagni, S., and Pirrelli, V. (2007). Automatic extraction of semantics in law documents. In *Proceedings of the V Legislative XML Workshop*, pages 253–266.

W3C OWL Working Group. (2009). OWL 2 web ontology language document overview. Technical report, W3C, October. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.

W3C OWL Working Group. (2012). OWL 2 web ontology language document overview (second edition). Technical report, W3C, December. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

CoUSBi: A Structured and Visualized Legal Corpus of US State Bills

Aikaterini-Lida Kalouli*, Leo Vrana*, Vigile Marie Fabella‡, Luna Bellani‡ and Annette Hautli-Janisz*

*Department of Linguistics ‡Department of Economics
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

This paper reports on an approach to automatically transform semi-structured and public databases of US state-level legislative bills into a structured, legal corpus, namely the Corpus of US Bills (CoUSBi). Our work has resulted in a methodology and a corpus that makes this data usable for natural language processing applications. It thus also lays important groundwork for work in the social sciences, particularly in the fields of political science and economics where there is a growing interest in the relationship between legislative policy-making and economic behavior. Against the backdrop of eventually contributing to a Legal Knowledge Graph, the paper shows that the corpus we provide already fulfills the requirements to be connected to other resources: We automatically extract correspondences between individual state bills and model bills from independent organizations, generating interesting insights into the legislative process. We furthermore use NEREx, a Visual Analytics framework, that allows us to capture important content of the bills at a glance.

Keywords: Resource development, US state bills, model bills, Visual Analytics

1. Introduction

As digitalization becomes increasingly infused throughout all aspects of society, it becomes even more important to make the legal domain and in particular the legislative process more accessible to the public. As a consequence, legislative bodies increasingly make information available online and users are faced with a flood of information, often unconnected to other relevant information and presented in a way that is not conducive to easy reference. This development is a classic case in which natural language processing (NLP) applications can be of great help. By way of structuring information from the legal domain in a certain way we can then design automatic systems that can shed light on aspects of legislative reality, e.g. answer specific queries in question-answering systems or in search engines, summarize legal documents, compare different versions of the same document and link related documents with each other.

Creating these structured resources across languages, legal traditions and types of legislative text involves standardizing information through interchange formats, for example as done with MetaLex (European Committee for Standardization, 2010), an interchange format for sources of law. For a general application of these standards, however, on the one hand a challenge lies in collecting documents from different kinds of sources and converting them into a standardized format. On the other hand, this format must also be common enough to be accessible for the communities involved, in particular NLP. With the aim of eventually creating a Legal Knowledge Graph, any approach faces questions about which key components need to be encoded to make specific types of legal text usefully accessible for further research.

In this paper we report on work that addresses the gap between legal text in the wild and a structured resource which can ultimately serve as input to a Legal Knowledge Graph. To this end, we collected all enacted, education-related bills of the US states North Carolina and New Mexico in the years 2007 to 2015. We automatically extract key information from the bills and convert them to structured documents according to the TEI standard (2017),[1] a common text encoding standard in the NLP community. We then link those bills to model bills of ALEC, the American Legislative Exchange Council,[2] established in 1975 (Hertel-Fernandez, 2014), which is a nonprofit organization in the US that drafts model state-level legislation.[3] In particular, we track down parts of the enacted bills that are also found in the model bills and we mark them accordingly within the structured documents. With this step we aim at detecting potential factors that drive a decision or the passage of a law. We also make a suggestion as to how to present the content of the bill texts: Using NEREx (El-Assady et al., 2017), a framework from Visual Analytics, we shed light on named entities and other important content words relevant in a particular bill.

The paper proceeds as follows: Section 2. presents related work and Section 3. describes our newly developed corpus CoUSBi (Corpus of US Bills) and discusses the challenges related to its development. In Section 4. we show how the standardized encoding of the bills allows us to make reference to other information, in particular the model bills of ALEC. Section 5. presents the visualization that we propose to use for displaying content information in the Legal Knowledge Graph. The paper closes with a discussion in Section 6..

2. Relevant work

Text mining in the legal domain is as varied as the type of data underlying it, ranging from extracting and analyzing arguments in legal cases (Moens et al., 2007; Wyner et al., 2010) to summarizing legal documents (Farzindar and Lapalme, 2004; Grover et al., 2003; Galgani et al., 2012) and constructing knowledge resources (Francesconi et al., 2010; Ajani et al., 2010, inter alia). Other efforts mine

[1] http://www.tei-c.org/index.xml
[2] https://www.alec.org/
[3] Available under https://www.alec.org/

legal terms (Pala et al., 2010; Surdeanu et al., 2010) or named entities (Quaresma and Goncalves, 2010; Dozier et al., 2010) in legal text. Another strand of research is concerned with automatic reasoning on legal text, bridging the gap between law and artificial intelligence (Hollatz, 1999; Bench-Capron and Sartor, 2003, among many others).

With respect to standardizing sources of law, the MetaLex initiative (European Committee for Standardization, 2010) has been at the forefront of providing an XML-based interchange format, with for example all Dutch regulations published in this format. In the humanities, the Text Encoding Initiative (TEI) standard is widely used to encode a wide variety of textual data.

With respect to these previous approaches, our work touches upon different aspects. Firstly, we create a structured resource from semi-structured US state bills and discuss the key characteristics that need to be encoded for this type of legislative data. The TEI standard we employ for this effort allows us to link information across different sources e.g. across the model legislation, a prerequisite for eventually contributing information to the knowledge graph. Lastly, the visualization tool can offer us insights in the content and relations of the corpus provided.

3. CoUSBi

3.1. Data collection

For now, CoUSBi consists of all enacted, legislative bills related to education between 2007 and 2015 from two US states, namely North Carolina and New Mexico. Both states offer their bills in a semi-structured and machine-readable format (in contrast to other states which only give the bill text as image or pdf). Creating a corpus of the enacted education-related bills (the rejected bills were irrelevant for the social science aims of the project), turned out to be difficult because such filtering was not catered for. We therefore invested a substantial amount of manual work in extracting the IDs of all enacted bills and then used $crawler4j$, an open source Java web Crawler,[4] to scrape the bills automatically. For now, the creation of the corpus depends on the painstaking task of HTML scraping which can be hard to maintain on a long-term. For the future, a systematic effort could create an API through which the states can directly deliver the bills. This is not meant to be extra work for the states: the present forming of the HTML structure also requires time to split the information of the bill in the corresponding HTML elements. An API could be a more user-friendly way of submitting this information.

The resulting resource consists of a total of 2,599 bills, with the actual text of each bill having an average of 3,257 tokens. We also include the different versions of the bills before their enactment and mark them accordingly in the file name (e.g. 'v1' for version 1). The bills have an average of 3.9 versions, with a maximum of 14 versions. The entire corpus is made available under `https://github.com/kkalouli/CoUSBi`.

[4]Available under `https://github.com/yasserg/crawler4j`

3.2. Encoding in TEI

CoUSBi is encoded in XML according to the TEI standard, a format widely used in the NLP community. Although TEI has not been designed specifically for legal text (in contrast to the aims of the MetaLex initiative, for example), it proves capable of handling the legislative bill data well, both with respect to the metadata and the actual text of the bill. For now we restrict ourselves to encoding the resource in the TEI standard, however a conversion to the MetaLex standard should be unproblematic.

As can be seen in the simplified XML overview in Figure 1, we need the following TEI elements to encode metadata and content structure in TEI (the full XML schema can be found within the resource): The header of the TEI header contains the element `fileDesc` which itself has three mandatory elements (`titleStmt`, `publicationStmt` and `sourceDesc`) and one optional element (`editionStmt`). Each of those elements contains a series of mandatory and optional subelements.

For encoding the body of the bill document, we use the `body` element with different subelements that specify the particular structure of the document. All in all, the following elements are included in each bill document:

- the short title of the bill (element: `<title>`)
- the authors of the bill: the representatives who took part in the writing process (element: `<author>`)
- the edition of the bill (element: `<edition>`)
- the publication place: the state the bill was presented in (element: `<pubPlace>`)
- the ID number: a combination of the state, year and bill number of the bill, e.g. NM-2013-S039 stands for the Senate Bill (S) with the number 039 of the state New Mexico (NM) and the year 2013 (element: `<idno>`)
- the source link: the url from which the bill originates (element: `<bibl>`)
- a short abstract which gives information on what the bill is about (element: `<head>` of the element `<div1 type="abstract">`)
- the text of the bill itself, separated into sections and paragraphs, according to the original sections and paragraphs
- further highlighting for underlined and strike-through parts of the bills: some bills have various versions (editions) and therefore some parts of them are either underlined to represent new parts or struck-through to represent parts that were removed from a later version. Since this information is important for the history of a bill, it is preserved and encoded in the TEI format.
- extra annotation whether a specific part of the bill is identical to a passage from a model bill (element: `<cit>` with subelements `<quote>` to hold the identical passage and `<bibl>` to hold the ID and section of the model bill it is identical to).

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader xml:lang="en">
    <fileDesc>
      <titleStmt>
        <title xml:id="NC-2007-H15">Txtbks Assignmts on Short-Term Suspension.</title>
        <author>Representatives Glazier, E. Warren, Parmon, and  Johnson (Primary Sponsors). </author>
      </titleStmt>
      <editionStmt>
        <edition>v0</edition>
      </editionStmt>
      <publicationStmt>
        <pubPlace>North Carolina</pubPlace>
        <idno>NC-2007-H15</idno>
        <date>2007</date>
      </publicationStmt>
      <sourceDesc>
        <bibl>www.ncleg.net\Sessions\2007\Bills\House\HTML\H15v0.html</bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text xml:lang="en">
    <front>
      <div1 type="abstract">
        <head>A bill to be entitled an act to implement a recommendation of the house select committee on the education of
        students with disabilities to allow students placed on short-term suspensions to take their textbooks home for the
        duration of the short-term suspension and to have access to homework assignments.</head>
      </div1>
    </front>
    <body>
      <div1>
        <opener>
          <salute>The <term type="NE" subtype="ORGANIZATION"> General Assembly of North Carolina </term> enacts:</salute>
        </opener>
        <div2 type="section" n="1">
          <head>SECTION 1. G.S.</head>
          <p>(b)The principal of a school, or his delegate, shall have authority to suspend for a period of <term type="NE" subtype="DURATION">
          10 days </term> or less any student who willfully violates policies of conduct established by the local board of
          <hi rend="strikethrough">education: Provided, that a</hi><hi rend="underlined">education.</hi>student suspended
          <hi rend="strikethrough">pursuant to</hi><hi rend="underlined">under</hi>this subsection shall be provided<hi rend="strikethrough">
          an opportunity to take any <term type="NE" subtype="SET"> quarterly </term>, semester or grading period examinations missed during the
          suspension period.</hi><hi rend="underlined">all of the following:</hi></p>
          <p>
            <hi rend="underlined">(1)</hi>
            <hi rend="underlined">The opportunity to take textbooks home for the duration of the suspension.</hi>
          </p>
          <p>
            <hi rend="underlined">(2)</hi>
            <hi rend="underlined">The right to inquire about homework assignments for the duration of the suspension.</hi>
          </p>
          <p>
            <hi rend="underlined">(3)</hi>
            <hi rend="underlined">The opportunity to take any <term type="NE" subtype="SET"> quarterly </term>, semester, or grading period examinations missed during the suspension period.</hi>
          </p>
        </div2>
        <div2 type="section" n="2">
          <head>SECTION 2. This act is effective when it becomes law.</head>
        </div2>
      </div1>
    </body>
  </text>
</TEI>
```

Figure 1: A simplified structure of the TEI-formatted bill of North Carolina H15 (version 0).

3.3. Automatic conversion

The conversion from HTML into the TEI XML format is done automatically with a rule-based system designed on the basis of the scraped HTML pages. The structure of the original webpages differs markedly across states, therefore each state requires its own conversion script. Concerning metainformation, we benefit from the fact that the file name of each bill consistently encodes the date, the bill number, the bill version and the state. Those are straightforwardly converted into their corresponding TEI elements. For the other TEI header elements, namely the representatives of each bill, the short title and the abstract, we use patterns in the bill text to extract the information. For example, in the North Carolina bills the representatives' names can be extracted by looking for the lexical pattern *Representative* or *Sponsor*, while in the New Mexico bills the information is encoded via the pattern *Introduced by*.

The body of each TEI document corresponds to the main body of the original bill. For transforming the paragraph elements of the HTML files, we apply straightforward conversion of the HTML element to the TEI element. So, excerpt (1) from a North Carolina HTML is converted to the TEI paragraph in (2) by moving the `<p>` element one-to-one and converting the `<s>` and `<u>` elements to the TEI elements `<hi rend="strikethrough">` and `<hi rend="underlined">`, respectively.

In order to include information whether parts of the text correspond to model bills of ALEC (for more information see Section 4.) we add the `<cit>` element that captures the matching parts.

(1)
```
<p class=amargin1>(b) The principal of a school, or his delegate, shall have authority to suspend for a period of 10 days or less any student who willfully violates policies of conduct established by the local board of <s>education: Provided, that a </s><u>education. A </u>student suspended <s>pursuant to </s><u>under </u>this subsection shall be provided <s>an opportunity to take any quarterly, semester or grading period examinations missed during the suspension period.</s><u>all of the following:</u></p>
```

(2)
```
<p>(b)The principal of a school, or his delegate, shall have authority to suspend for a period of <term type="NE" subtype="DURATION"> 10 days </term> or less any student who willfully violates policies of conduct established by the local board of<hi rend="strikethrough">education: Provided, that a</hi><hi rend="underlined">education. </hi>student suspended<hi
```

```
rend="strikethrough">pursuant
to</hi> <hi rend="underlined">under
</hi>this subsection
shall be provided<hi
rend="strikethrough">an
opportunity to take any <term
type="NE" subtype="SET"> quarterly
</term>, semester or grading
period examinations missed during
the suspension period.</hi><hi
rend="underlined">all of the
following:</hi></p>
```

3.4. Named entity annotation

For tagging named entities, we use the Stanford Named Entity Recognizer (Finkel et al., 2005), and conclude that further model training is not necessary. The text of each bill section was fed into the recognizer, and a script parsed the output to add XML tags around the entities identified. Tags for *NUMBER* were disregarded, as their extremely high frequency led to so many labels in section headings as to be no longer of use. In excerpt (2) above we can see how the named-entity *10 days* is marked with the element <term> of the type *NE* (for Named-Entity) and the subtype *DURATION*.

3.5. Challenges

The main challenge in creating CoUSBi was to convert inconsistent information of the source into a standardized format. There were several types of inconsistencies. Firstly, there were formatting inconsistencies in the HTML encoding of bills within one state. For example, some bills would be encoded in CSS, while other bills contained a mixture of CSS and HTML, but encoding the same information.

A second group of inconsistencies was the incomplete information in the source. This means that not all bills within one state encoded the same kind of information, i.e. they did not include the same document elements. This inconsistency is tightly bound to a third one, namely the misplacement of some of the relevant information. Many of the bill documents contained the relevant information but not always in the same position within the document. This meant, for example, that although in most of the bills the date information was found after the title, in some of them we had to look for the date elsewhere. We also observed that many smaller details relevant for the task were not consistent, e.g. the use of capital or small letters for specific elements.

As a consequence of these issues and also because some of them were so profound, we have so far not converted the bills of West Virginia — an additional state on which we had started working — into the TEI format, as this requires further extensive manual effort. As it is, conversion of the bills from the other two states has already been very time-consuming. Although the issues mentioned can be solved with relatively simple, additional rules, e.g. for the inconsistency of small-capital letters we can use case-insensitive rules, it is tedious to make sure that all such inconsistencies have been traced and handled. Although we cannot provide a formal evaluation as to the completeness and accuracy of the resource, we believe that our repeated attempts to detect and handle all inconsistencies have paid off in a way that there is no missing information.

4. Linking information

One of the defining characteristics of knowledge graphs is that they link information from different sources. Despite a comparatively preliminary size and coverage of CoUSBi, we are nevertheless able to make interesting connections and comparisons and show that even preliminary investigations shed light on legislative processes as a whole.

In the US, many different organizations produce and distribute draft legislation which can be adopted and adapted by the concerned authorities. One such organization is the American Legislative Exchange Council (ALEC), with members including politicians as well as corporate representatives. Together they produce model bills that can then be directly introduced for debate in state legislatures (Hertel-Fernandez, 2014). Some states such as Arizona, Wisconsin, Colorado, Michigan, New Hampshire, and Maine make heavy use of the ALEC model bills (Rizzo, 2012). Approximately 200 ALEC bills become law each year (Greenblatt, 2003). As part of our work on CoUSBi, we investigated whether any bills introduced in state legislatures include influences from ALEC bills.

Preparing the ALEC bills All education model bills were scraped from the ALEC website, where they are freely available. There were 74 education bills in total posted on the ALEC website at the time of access, with dates ranging from 1995 to 2017. Some model bills did not include a date. It is not immediately clear in every case whether these dates reflect the online publishing date, or the date they were originally drafted. We also accept that this time frame both predates and extends beyond the dates of the bills collected in our corpus. However, model legislation that is several years old is by no means past its shelf life, and legislation introduced by a state may be then copied and distributed as model legislation, a relationship which would also be of interest.

Once scraped, each piece of model legislation was converted to the TEI standard consistent with the elements listed above for state legislation. Although most of this information is not available for the model bills, keeping these elements consistent will facilitate future analysis. Further, automatic annotation of bill sections was possible, providing important reference points for comparison with state bills. In all, this process produced a second smaller TEI corpus of ALEC education bills.

Linking ALEC bills and state bills In order to determine sections of ALEC bills which provided relevant matches to passages of bills introduced in the state legislatures, simple 15-grams from the text of each piece of model legislation were searched for in the text coming from the bills. In order to aid matching, the text of the bills was stripped of punctuation, and case was ignored. Further matching was able to determine which passages of the bills matched the passages in model legislation, which could then be automatically annotated. The annotation uses the <cit> element of the TEI format, which features the

North Carolina	**ALEC**
The Founding Principles Act, H588v0 & v1 (2015)	The Civic Literacy Act (No date)
Whereas, the adoption of the Declaration of Independence in 1776 and the signing of the United States Constitution in 1787 were seminal events in the history of the United States, the Declaration of Independence providing the philosophical foundation on which the nation rests, and the Constitution of the United States providing its structure of government; and Whereas, the Federalist Papers embody the most eloquent and forceful argument made in support of the adoption of our republican form of government; and Whereas, these documents, along with the writings of the Founders, stand as the foundation of our form of democracy, providing at the same time the **touchstone** of our national identity and the vehicle for orderly growth and change; and Whereas, these Founding Documents established a set of principles, known as the Founders' Principles, which are the heart and soul of a government for a free society; and Whereas, these principles enabled a group of 13 colonies to become the greatest and most powerful nation on earth in a relatively short period of time; and Whereas, most Americans do not know about nor understand the timely and timeless importance of these principles to our form of government and to their current lives; and Whereas, the survival of the republic requires that our nation's children, the future guardians of its heritage and participants in its governance, have a **clear understanding** of these principles and the importance of their preservation;	(A) The adoption of the Declaration of Independence in 1776 and the signing of the United States Constitution were principal events in the history of the United States, the Declaration of Independence providing the philosophical foundation on which this nation rests and the Constitution of the United States providing its structure of government. (B) The Federalist Papers embody the most eloquent and forceful argument made in support of the adoption of our republican form of government. (C) These documents stand as the foundation of our form of democracy providing at the same time the **basis** of our national identity and the vehicle for orderly growth and change. (D) Many Americans lack even the most basic knowledge and understanding of the history of our nation and the principles set forth in the Declaration of Independence, codified in the Constitution and defended in the Federalist Papers. (E) The survival of the Republic requires that our nation's children, the future guardians of its heritage and participants in its governance, have a **firm knowledge and understanding** of its principles and history.

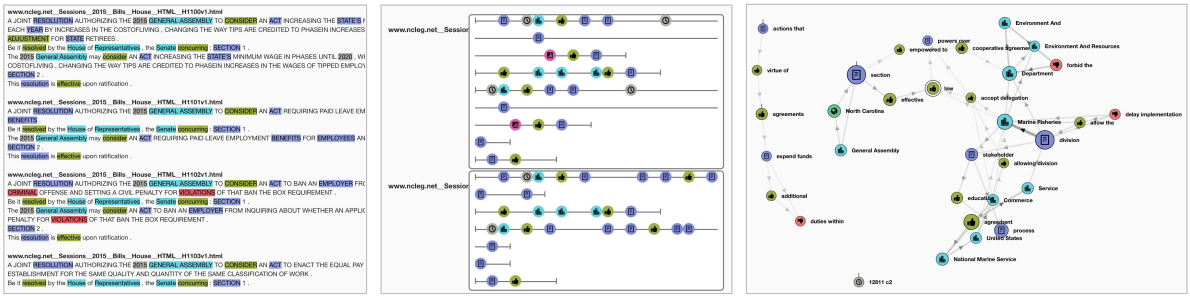
Figure 2: In the example above, green highlighted sections are identical, and closer reading reveals other sections, highlighted in teal, that are highly similar. Bold words indicate differences in otherwise similar passages.

subelements <quote> and <bibl>. The <quote> element contains the passage that is taken from the model bill and the <bibl> element specifies the ID and the exact section of the matching model bill.

Results The results of this analysis found that 13 non-overlapping verbatim spans of 15 words or longer from model bills were also found in North Carolina state education bills during this time period, and 10 portions were found in New Mexico's state bills. These spans ranged from 17 to 36 words in North Carolina's bills and 16 to 36 words in New Mexico's bills. The length of the n-gram threshold helped to ensure the retrieval of relevant similarities. Some passages seemed to offer formulaic speech for a definition and contained no resemblance in language or structure in the surrounding context. Other passages revealed that the sections preceding and/or following the verbatim passage contained multiple slight alterations that did not alter the meaning of the text, but which did prevent our method from identifying it as an extended block of verbatim text.

Such methods "will never replace careful and close reading of texts" (Grimmer and Stewart, 2013), and indeed this method can be of most utility in flagging sections for further examination, an example of which is shown in Figure 1. Examining the dates associated with the documents shows that only a small proportion of the matching state bills were introduced after the model legislation's reported date on ALEC's website (4 of 23 passages). However, this does not preclude the possibility that the passages in the state legislation were written by ALEC, or another organization. Previous study into ALEC has depended on internally leaked documents, as the group strives to keep a low profile (Hertel-Fernandez, 2014), and it is known to operate by distributing legislation to lawmakers without making this process public. Furthermore, the extent to which this language is mirrored between sources suggests that there is some link between the source of these passages.

This type of analysis can be helpful in analyzing legislation to identify similarities between states and other entities, and how influence may manifest itself in shorter passages, even if the full bill does not completely adhere to the goals of model legislation. Other methods such as fuzzy string matching and Levenshtein edit distance calculations could be employed in order to find matches with subtle differences, and to gauge whether the similarities are meaningful or coincidental.



(a) Text-Level View (b) Entity-Level View (c) Entity Graph

Figure 3: Named-Entity Relationship Explorer

5. Visualization

One of the core aims of encoding information in a knowledge graph is to present this information at a glance. In the case of a large amount of bills, a crucial component is to give the user an idea of the content of the bill, i.e. provide an intuitive overview of important concepts and entities that are covered in the bill. To this end we employ NEREx (El-Assady et al., 2017), a Visual Analytics framework for the analysis of different concepts and their relation in the utterances. Using Visual Analytics for this task is motivated by the challenge of dealing with large amounts of data, while at the same time providing the user with an interactive and exploratory access to the data (Keim et al., 2008).

The data is uploaded through a web interface and relevant named-entities and concepts from the text are categorized into ten classes: Persons, Geo-Locations, Organizations, Date-Time, Measuring Units, Measures and Context-Keywords. Using a perceptually preattentive visual encoding for these categories, the text is abstracted from the Text-Level View (Figure 3a) to the Entity-Level View (Figure 3b) to allow a high-level overview of the entity distribution across utterances.

For extracting relations in the text, the framework uses a tailored distance-restricted entity-relationship model, which relates two entities if they are present in the same sentence within a small distance window defined by a user-selected threshold. The concept map of the conversations can then be explored in the Entity Graph (Figure 3c). All views support a rich set of interactions, e.g., linking, brushing, selection, querying and interactive parameter adjustment.

The visualization supports the analyst in two ways: First, the *content of the bill* can be displayed with increasing abstraction, catering for different demands of the analyst (from close reading to distant reading). The Entity Graph gives an overview over highly relevant terms: The more saturated the colors of the arcs, the more frequent the nodes (i.e. entities/concepts), with the direction of the arc showing the order of the items.

For illustrative purposes, we use NEREx to display the content of only one bill, namely the 2015 bill S140 from the North Carolina Senate, which authorizes the town of Lake Santeetlah to levy an occupancy tax. Figure 4a shows the Entity Graph for the complete bill, the subgraphs in Figures 4b and 4c zoom in on the upper and lower middle part of the overall graph, respectively. In subgraph 1, the terms ‘Tourism’ and ‘Authority’ are at the center, with the former connected to the bigrams ‘Town ,’ and ‘Town .’, which in turn co-occur with ‘attract tourists’ and ‘expenditures’. Consulting the Text-Level View shows that the bill specifies the way the Tourism Authority spends the tax revenue: By funding tourism-related expenditures and support tourism-related activities. Subgraph 2 shows all cities belonging to the Lake Santeetlah Council that levy an occupancy tax – these cities are plotted on the canvas according to their geolocations. These examples show that an analyst can use the Entity Graph to get an immediate overview of the content of the bill, with the Text-Level View allowing for a more detailed investigation of the actual text.

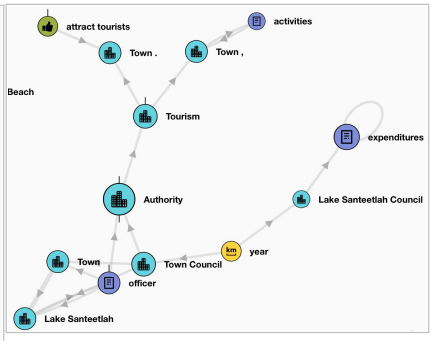
The visualization is also important for *resource validation*: As we were going through the visualization of individual bills, the Entity-Level View showed that in a small number of cases the bill content was repeated, based on inconsistencies in the source. In other cases, the bill text was blank, also due to erroneous HTML encoding in the source. These errors were then manually corrected in the corpus.

6. Conclusion and future work

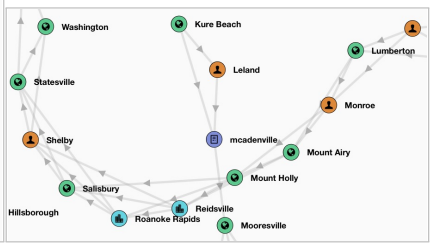
This paper reported on an approach of mining legal data in the wild and the requirements, challenges and potentials that go with it. Such a resource can be part of the Legal Knowledge Graph and can be used by professionals in the legal domain to examine and monitor the lawmaking process, from influence, to drafting, to editing, and the passage into law. Creating this type of structured resource also lays the groundwork for other research in the social sciences. One concrete application in political economy uses education bills to determine how elected officials respond to the preferences of their constituency. When the school performance of students in their electoral district weakens, do they respond by authoring a particular kind of education bill? How is the support for a bill in the legislature affected by additions or removals of certain clauses? To what extent do bill authors make policy tradeoffs between the preferences of his constituency and the preferences of the opposition? Such questions can only be answered using information from corpora such as CoUSBi.

Due to its consistent encoding in TEI, CoUSBi can also be utilized as-is for further syntactic and semantic parsing or can be indexed and used for direct query processing. It is also — to the best of our knowledge — the first attempt to automatically compare US bills to model legislation. The identical language in passages suggests a connection which

(a) Content of bill S140 from North Carolina



(b) Subgraph 1 of Figure 4a



(c) Subgraph 2 of Figure 4a

merits further analysis. Examining content on this level requires an extremely labor-intensive effort for human readers and the automatic method presented in this paper illustrates just one technique which could prove valuable to this end. As the corpus is expanded to include further legislatures and more model legislation, this type of research could be expanded, providing the public with its own measure of such organizations. Other research into paraphrasing, as well as other text matching methods could help to identify corresponding sections between model legislation and bills proposed in the states. Thus far, this information has only been available through painstaking reading through several bills. While this is only a first step, this kind of monitoring becomes possible with the advent of standardized and openly accessible legislative corpora.

Besides our goal to include more bills across states and topics, we also aim at implementing a framework that automatically compares different versions of the same bill and offers some insights on the types of changes between different versions. This will include, for example, an at-a-glance overview of sections withdrawn from or added to bills or highlight those that were only slightly modified. We also see a potential in applying topic-modeling techniques to the corpus in order to annotate individual bills with their key topics. We would additionally like to make use of the full potential of NERs by linking entities of different bills to each other and to the LOD [5] cloud, in this way making a vast amount of knowledge accessible. We furthermore consider the conversion of the TEI-formatted resource into the MetaLex standard, also to facilitate a comparison of legislation across languages and traditions.

[5]Linking Open Data, available under `http://lod-cloud.net/` and `http://linkeddata.org/`

7. Bibliographic References

Ajani, G., Boella, G., Lesmo, L., Martin, M., Masszei, A., Radicioni, D. P., and Rossi, P. (2010). Multilevel legal ontologies. In *Semantic Processing of Legal Texts*, pages 136–156. Springer: Berlin Heidelberg.

Bench-Capron, T. and Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1-2):97–143.

Consortium, T. (2017). Tei p5: Guidelines for electronic text encoding and interchange.

Dozier, C., Kandadadi, R., Light, M., Vachher, A., Veeramachanenin, S., and Wudali, R. (2010). Named entity recognition and resolution in legal text. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 27–43. Springer: Berlin Heidelberg.

El-Assady, M., Sevastjanova, R., Gipp, B., Keim, D., and Collins, C. (2017). Nerex : Named-entity relationship exploration in multi-party conversations. In Jeffrey Heer, editor, *EuroVis 2017 Eurographics / IEEE VGTC Conference on Visualization 2017*, number 36,3 in Computer Graphics Forum, pages 213–225.

Farzindar, A. and Lapalme, G. (2004). Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out: Proceedings of the ACL 2004 Workshop*, pages 27–34.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics ACL2005*, pages 363–370. Association for Computational Linguistics.

Francesconi, E., Montemagni, S., Peters, W., and Tiscornia, D. (2010). Integrating a bottom-up and top-down methodology for building semantic resources for the

multilingual legal domain. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 95–121. Springer: Berlin Heidelberg.

Galgani, F., Compton, P., and Hoffmann, A. (2012). Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012*, pages 115–123.

Greenblatt, A. (2003). Governing.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Grover, C., Hachey, B., and Korycinski, C. (2003). Summarising legal texts: Sentential tense and argumentative roles. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 33–40.

Hertel-Fernandez, A. (2014). Who passes business's "model bills"? policy capacity and corporate influence in us state politics. *Perspectives on Politics*, 12(3):582–602.

Hollatz, J. (1999). Analogy making in legal reasoning with neural networks and fuzzy logic. *Artificial Intelligence and Law*, 7(2):289–301.

Keim, D., Andrienko, G., Fekete, J.-D., Górg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In A. Kerren, et al., editors, *Information Visualization*, pages 154–175. Springer, Berlin.

(2010). Metalex (open xml interchange format for legal and legislative resources).

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law. Association for Computing Machinery*, pages 225–230.

Pala, K., Rychl'y, P., and Smerk, P. (2010). Automatic identification of legal terms in czech law texts. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 83–94. Springer: Berlin Heidelberg.

Quaresma, P. and Goncalves, T. (2010). Using linguistic information and machine learning techniques to identify entities from juridicial documents. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 44–59. Springer: Berlin Heidelberg.

Rizzo, S. (2012). Some of christie's biggest bills match model legislation from d.c. group called alec.

Surdeanu, M., Nallapti, R., and Manning, C. (2010). Legal claim identification: Information extraction with hierarchically labelled data. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Text, co-located with LREC 2010*, pages 22–29.

Wyner, A., Mochales, R., Moens, M.-F., and Milward, D. (2010). Approaches to text mining arguments from legal cases. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 60–79. Springer Berlin Heidelberg, 01.

Legal Document Similarity using Triples Extracted from Unstructured Text

Akshay Minocha, Navjyoti Singh
IIIT-Hyderabad
Hyderabad, India
akshay.minocha@research.iiit.ac.in, navjyoti@iiit.ac.in

Abstract

This research is an attempt to address the complication of bridging the gap between the traditional systems and the future of legal systems. We discuss one of the facets of the process of legal understanding and decision-making in the court of law, as well as aim to increase general public comprehension on the topic of constitutional importance. We focus on a selected list of documents gathered through citation network analysis, and using the knowledge of the Sections in the Income Tax Act of India which govern them; after processing the proceedings identified, through the proposed technique. The resulting triples are used to evaluate the similarity of such legal documents.

Keywords: legal information retrieval, document similarity, ontology, knowledge graph, semantic triples

1. Introduction

With the development and countrywide acceptance of internet-centric applications which fall under the category of e-governance in India; the legal domain is one area of interest which deserves great mention. The availability of reasonable sources and legal data, help in the building of practical and assistive applications which would be of great use to the legal experts. To a domain practitioner, there are other detailed applications such as - document classification, legal knowledge discovery, legal information retrieval, predictive mechanisms, and so on.

Many efforts have been made in this field to facilitate faster and better legal help to legal practitioners, advocates, researchers and the non-domain people. Although the accessible legal resources in India have been recently made available and there is a lot of data which can be used to increase the efficiency of these services, yet this information by large remains unstructured. In our research, we aim to look at cases which belong to a specific category of cases adhering to finance and income tax. We have generated an ontology of the sub-domain of the legal area and try to align the cases which cite these Sections of the Act according to the triples made by the technique proposed in Section 3 . The similarity of these cases is evaluated based on a thematic scheme, and we then discuss the results in Section 5 where a complete state of the situation and further steps are mentioned.

2. Related Work

In (Kumar et al., 2011), (Kumar et al., 2013) the authors use statistical measures and connective properties in text to predict the similarity of legal judgments, and on the other front (Saravanan et al., 2009), (Saravanan et al., 2006), talk about a novel method of legal document summarizing and effective retrieval by suggesting that we approach the problem with an ontological perspective. For legal information retrieval, it is the objective to manufacture an intuitive data space to consequently outline content information to an adjustable ontology. Legal Ontological enquiry has been inspired by the work done by LKIF (Hoekstra et al., 2007), (Breuker et al., 2007) where they also come up with a legal information interchange format along with an ontology of basic legal concepts in Italian law. The work is really inspirational in terms of providing a movement towards a knowledge representation formalism in the legal domain.

A triple as the name suggests is a combination of three different sets of words, an atomic form of information which provides semantics to the situation or in our case the legal text in hand. Just put into words it is a subject-predicate-object expression. Just like we have specific grammar while writing computer programs we have to find out a way in which we can simplify phrases and sentences into a more machine-readable format. A sentence can be broken down into multiple triples according to its complexity. Triples are one of the many ways in which information from a judgment is presented in a less complicated manner with fewer relevant words.

Understanding the relational facts from understandable content has for quite some time been of enthusiasm for data extraction research. The critical issue is to adjust the exchange off between high precision, recall, and adaptability. With the rise of the Semantic Web and various ontologies, information combination has turned into an extra test. There has been a lot of research on semi-supervised strategies utilizing bootstrapping methods together with beginning seed relations to make extraction designs. Unsupervised methodologies have contributed to work in the legal domain by not requiring hand-tagged information. These methodologies have addressed efficiently versatility and accuracy factors when connected on web-scale corpora. A system like LODifier (Augenstein et al., 2012) is a cornerstone in the achievements towards triple extraction research. Our data is not as much tagged and linked to entities so that it can easily be mapped onto a very well developed knowledge base, as the (Exner and Nugues, 2012), we connect the extracted entities in an unsupervised way which in turn would bring form and structure to the legal domain knowledge base.

3. Methodology

3.1. Dataset

The Income Tax Act of India was authorized in the year 1961 and is the statute under which everything identified with tax collection is recorded. The Act incorporates levy, collection, organization, and recuperation of wage assessments. The act represents a constitutional reference for people seeking support from a consolidated set of rules identified with tax collection in the nation. Organized into over 23 chapters and many schedules the act covers a great deal of the laws which are to be followed by individuals, firms, partnership firms about their dealings and their functioning. Due to its large breadth, we wanted to cover a specific part of the act which would comprise of knowledge which is in some way self-sufficient. The method of identifying this group of Sections within the Act was to generate by scraping[1] all the cases that cite the individual parts of the act and then narrow down to the division which shows the highest coverage in terms of the ratio of cases which only deal with this part of the act. The citation connections between the legal documents were made in a similar way as implemented to find graph connective measures for various network properties in the legal domain (Minocha et al., 2015). This grouping of cases, would ensure an investigation on an independent group of the act where most of the cases can be categorized into and belong within the sub-domain of the legal knowledge. The reason for this was to come up with an ontology which is tending to complete on its own with little dependencies on other parts of the act. We chose the part of the Income Tax Act which deals with 'Changes in constitution, succession, and dissolution of firms and partnerships'. The Sections of the act which were of interest are Section 187, 188, 188A, 189, and 189A. The number of such cases until the day of investigation was close to 80, and this number also seemed perfect to experiment with the methodology discussed later on in the paper.

3.2. System Architecture

Figure 1, explains the different modules that are involved in the work-flow of the system to generate some tagging and triples for the documents so that they can be compared against each other for similarity.

- **Headnote extraction**: The legal proceedings and documents available usually have the facts and a summary related to the case in the initial part of the document. In most of the documents such is the case, as the court proceedings first comprise of the known and acknowledged facts that have been put forward as the basis of the case. Extracting this part is crucial for us since we do not want the remaining text which would include discussions related to different cited cases, references, and opinions that might not be facts yet. Such data in triples can be conflicting, and hence headnote is extracted heuristically from the proceedings for our research.

[1] https://indiankanoon.org/

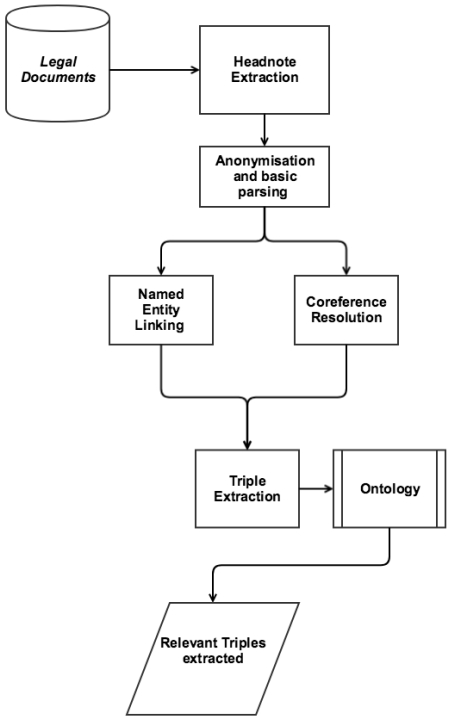


Figure 1: Triple Extraction from Legal Text

- **Anonymisation**: In this stage we try to anonymise the names of the partners and the firms so that more overlapping structures can be made while linking the triples.
- **Named Entity Linking and Co-reference Resolution**: Legal texts are lengthy, and constructing rules to extract triples becomes increasingly difficult, resulting in either very lengthy relations or issues in correct noun phrasal entity inclusion. To tackle this problem and to make the process efficient, we perform these text pre-processing tasks to help in obtaining relevant triples.
- **Triple Extraction**: For the triple extraction we use OpenIE (Fader et al., 2011),(Etzioni et al., 2008), a confidence score is obtained along with the extracted triple. We implement an instance of OpenIE in our work which in a single pass extracts a large set of relational tuples from the data. OpenIE does not require any human intervention in labelling or input
- **Ontology Mapping**: In this phase we map the triples to the common terms and actors which have been identified by describing the ontology of the legal domain in question, by doing this we create a set of triples which would have high overlap when the input legal cases are similar, due to their standardised nature.

4. Evaluation

To categorize or find similar legal proceedings which handle intricacies related to legal entities in a conceptual way.

We provided more than 80 random pairs of judgments from our dataset to three legal experts and asked them to rank these documents concerning similarity in the range of 1 to 10 (Raghav et al., 2015). Information about the dataset was given with the evaluation exercise. A similarity score of 10 would mean that the documents have a great deal in common and can be treated as a reliable reference by a legal practitioner while preparing arguments. The score does not represent a binary classification because of the nature of the extent of similarities which is planned to be used in case of future experiments. Although, we would be using this in a binary form for our analysis at hand, details of which follow.

Inspired by the work done for LODifier (Augenstein et al., 2012), our similarity measure is based on the distance and overlap between similar nodes. A short path indicates more relevant semantic information.

In Section 3.1. we described how that gold data which elucidates the similarity function, concerning scores are annotated for similarity by professionals and legal practitioners. However for comparison with other metrics and the extensions with the proposed changes we will use story link detection test, which when used initially, analyzed the information where two randomly selected stories to discuss the same news topic (Augenstein et al., 2012).

We have a score computed for each approach which is termed as sim, and like setting a base threshold, we have a similar limit here for which the following classification holds -

$$class(d_p, \theta) = \begin{cases} positive, & \text{if } sim(d_p) \geq \theta \\ negative, & \text{otherwise} \end{cases}$$

According to the above equation, a document is said to be similar if it has a similarity of θ or above, $theta$ being our threshold for the assertion. The central statement here lies in computing the θ parameter for each experiment. According to the investigation, we would use cross-validation to split our dataset into test and train classes. In this case θ would predict the training set in each iteration of the tuning (let's say $k = 100$), as well as possible. The distance of θ would be such that it would maximize the number of similar pairs.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\sum_{d_p \in pos\_train} min(0, sim(d_p) - \theta)^2 + \sum_{d_p \in neg\_train} min(0, \theta - sim(d_p))^2 \right]$$

We can then over more tuning iterations predict a better and more accurate value of θ.

$$prosim_{k,Rel,f}(G_1, G_2) = \frac{\sum\limits_{\substack{a,b \in Rel(G_1) \\ <a,b> \in C_k(G_1) \cap C_k(G_2)}} f(l(a,b))}{\sum\limits_{\substack{a,b \in Rel(G_1) \\ <a,b> \in C_k(G_1)}} f(l(a,b))}$$

We use a measure called $proSim$ which translates to path relevance overlap similarity (Augenstein et al., 2012). When, $f(l) = 1$ this counts the number of paths irrespective of the length. We do this since unlike the other tasks the graphs generated from the *headnote* are not massive and hence very long and complicated paths are not encountered. Accordingly, we also select the graph with more number of nodes since if the documents are somewhat similar G_1 will absorb the facts conveyed by G_1.

5. Results

We chose to see a more additional correlation, in light of the most limited ways between similar documents. This mirrors our instinct that a more informative structural source catering to two documents means a striking semantic connection and a similar theme between them as well.

The other methods against which the evaluation task has been held is a cosine similarity model - a standard evaluation metric in the domain of corpus-based document evaluation, used as a baseline in many situations. However, the disadvantage of this metric are that this is somewhat based on a bag-of-words (BoW) model. Therefore, it does not take into account the position of the word in the text, semantics, and co-occurrences. Nonetheless, the results of this metric are not very poor because the documents belong to one domain and have similar kind of terminologies mentioned in them; however much complex rules and semantic relations as discussed are not captured which makes this metric not credible concerning finding similarity for legal cases.

Technique	**Accuracy**	**F1-Score**
Our Method	73.17%	0.807
L_{mod}	59.7%	0.718
Cosine Similarity	54.87%	0.53

Table 1: Results from techniques mentioned in Section 5

The other metric is to compare with similar research which had taken place in the Indian legal context and is important research regarding Legal Ontology-based inquiry (Saravanan et al., 2009). The original extract of the legal ontology as mentioned is modified to deliver better results, L_{mod}. The initial extract was a workaround of all legal cases; we reduced the acts to be the Sections and also introduced primary events such as death, retirement, penalty, etc., so that the results are somewhat comparable. The results show that our design technique for the comparable metric is promising, but since the whole idea of designing a specific ontology is to get better results, a modification to a particular use case cannot do justice to the original purpose. We did not choose metrics related to co-citation networks since the dataset has been designed keeping in mind the same systems and hence the comparisons would not hold proper meaning.

6. Conclusion and Future Work

In this particular research, the work is related to validating the concept of the ontology based triples, and the same helping in the assessment of similarity of legal documents.

The number of proceedings positively affected the results concerning differences, and assessing more reports by including more divisions to the Income Tax Act would only heartily approve of the technique in future. Some things that we would want to point out are that the court upholds the law in the best possible way, by the rules defined in the Sections and the corresponding ontology.

On analyzing further, we saw that some inefficiencies in the results were due to some facts being furnished later on in the proceeding after discussions, or that there were some facts which at the time of being provided initially are wrong which is then rectified in the text. With more triples, we can even generate a triple-store for the cases and a query based information retrieval mechanism can help in finding out proper precedent for the situation at hand. Even though there were cases which were complicated, there were also cases which were similar and redundant; this just reflects the inefficiencies of the administration and the lack of information about the law amongst the public.

A natural clarification for the execution of our ontology-based framework is that it gives a knowledge base which has an enormous accumulation of terms and its connections and other related components which are utilized for better upgrades of query terms. Likewise, our basic structure can be extended with the expansion of conditions by including new archives, and from different subdomains in the future course of time.

We would like to conclude by saying, that the results are promising, and more efficient ontological rules across a broader spectrum of legal norms along with more efficient triple alignment techniques can help us further, with not only better document similarity metrics, but also in terms of a legal knowledge graph with untapped potential in terms of applications aiming to find precedents, similar judgments and understanding legal constraints along the way.

7. Bibliographical References

Augenstein, I., Padó, S., and Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In *Extended Semantic Web Conference*, pages 210–224. Springer.

Breuker, J., Hoekstra, R., van den Berg, K., Rubino, R., Sartor, G., Palmirani, M., Wyner, A., Bench-Capon, T., et al. (2007). Owl ontology of basic legal concepts (lkif-core).

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Exner, P. and Nugues, P. (2012). Entity extraction: From unstructured text to dbpedia rdf triples. In *The Web of Linked Entities Workshop (WoLE 2012)*, pages 58–69. CEUR.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.

Hoekstra, R., Breuker, J., Di Bello, M., Boer, A., et al. (2007). The lkif core ontology of basic legal concepts. *LOAIT*, 321:43–63.

Kumar, S., Reddy, P. K., Reddy, V. B., and Singh, A. (2011). Similarity analysis of legal judgments. In *Proceedings of the Fourth Annual ACM Bangalore Conference*, page 17. ACM.

Kumar, S., Reddy, P. K., Reddy, V. B., and Suri, M. (2013). Finding similar legal judgements under common law system. In *International Workshop on Databases in Networked Information Systems*, pages 103–116. Springer.

Minocha, A., Singh, N., and Srivastava, A. (2015). Finding relevant indian judgments using dispersion of citation network. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1085–1088. ACM.

Raghav, K., Reddy, P. B., Reddy, V. B., and Reddy, P. K. (2015). Text and citations based cluster analysis of legal judgments. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 449–459. Springer.

Saravanan, M., Ravindran, B., and Raman, S. (2006). Improving legal document summarization using graphical models. *Frontiers in Artificial Intelligence and Applications*, 152:51.

Saravanan, M., Ravindran, B., and Raman, S. (2009). Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law*, 17(2):101–124.

Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel
Ontology Engineering Group, Universidad Politécnica de Madrid
{emontiel, vrodriguez}@fi.upm.es

Abstract

This position paper describes the motivation, objectives and expected results of the recently started European project Lynx (H2020-780602). Lynx aims to provide a set of smart services to assist companies in Europe with compliance needs. The novelty of the proposed compliance services is that they will be built on and exploit a graph of legal and regulatory data – from different jurisdictions and in different languages – duly integrated according to the Linked Data paradigm. The fact that the Legal Knowledge Graph will bring together data from different legal and regulatory traditions in several languages is one of the most challenging aspects that language technologies will help to overcome.

Keywords: compliance, Legal Knowledge Graph, Linked Data, multilingualism, language technologies

1. Motivation

SMEs (Small and Medium Enterprises) introducing a product or service in a new market face fundamental issues related to legal and regulatory aspects that affect different markets. Questions that might have to be solved before going international include (by means of example):

How does country X deal with this aspect regulated in my country by regulation R? Which consequences could that have for the product or service I would like to launch?

How is the technical question Q handled in country Y? How has it been understood by local courts?

Which standards should we implement before launching our products or services?

In order to help companies to deal with these and similar compliance issues, the Lynx solution will create a unique and novel knowledge base related to compliance integrating information from heterogeneous data and content sources in what we have termed the Legal Knowledge Graph (LKG). Since compliance has to do with "the conformance to a set of laws, regulations, policies, or best practices" (Silveira et al., 2010), deeply rooted in each country's traditions and which are mostly expressed in its own language, the integration task is a challenging one. In this sense, we believe that the combination of Semantic Web approaches, specifically the Linked Data paradigm, and language technologies will help us perform the integration phase. As a result, a set of services will be built upon the LKG to demonstrate the feasibility of the approach and solve some of the current business needs represented by three pilots on the following topics: a) data protection, b) regulation in the oil and gas industry, and c) labour law, as will be explained in Section 5.

Lynx is a European research project funded as an H2020 Innovation Action covering the topic ICT-14: Big Data PPP: *cross-sectorial and cross-lingual data integration and experimentation*. The project began in December 2017 and will run for three years.

2. Main Objectives

The main objective of Lynx is to facilitate compliance of SMEs in internationalisation processes by leveraging existing European legal and regulatory open data. This will be achieved by providing innovative legal business models by connecting national and international legislation, regulations, standards, case law, and best practices in the same or different languages. Building such a cross-language legal information system is aimed to offer a competitive advantage for European companies against others operating within a single jurisdiction. As such, these companies will manage to reduce the costs and efforts related to organising and monitoring legislation, regulations and sectorial good practices, and even incorporating new jurisdictions into their businesses.

As for the technical objectives, these can be classified into three: 1) to deliver domain-neutral common services to create and exploit the LKG (document annotation, interlinking, term extraction, smart search, translation, etc.), 2) to create a family of business-oriented applications to cover the particular needs of the user cases involved in the project, and 3) to provide a single entry point to interlinked legal information across jurisdictions and languages.

Beyond those business and technical objectives, Lynx also pursues societal objectives. Amongst others, providing European citizens better access to legal and regulatory information from multiple jurisdictions, and an easier involvement in legislative processes.

3. The Lynx Data Value Chain

The Lynx approach relies on a data value chain that consists of three main phases: data acquisition, data integration and data exploitation, as illustrated in Figure 1.

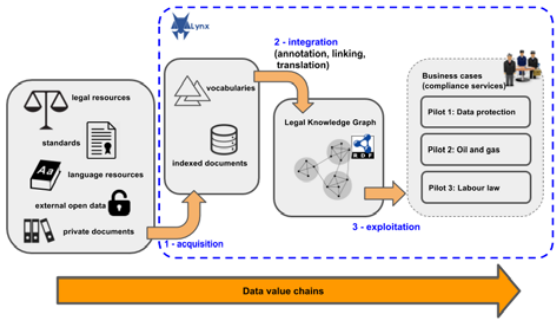


Figure 1 : Lynx Data value chain

Step 1. Acquisition: Regulatory data in a broad sense (legislation, jurisprudence, standards, and norms) from multiple jurisdictions and in different languages will be aggregated in the LKG.

The EU PSI Directive, in force since 2003 and amended in 2013, was followed by national developments in the Member States which unleashed huge amounts of high quality data in government data portals. Central to the domain of compliance are the European Data Portal[1] and the Eur-Lex[2] portal, maintained by the EU Publications Office, and hub of European Union law and other official public documents.

Additionally, a number of public and private-led initiatives are publishing online many valuable data sets and resources, openly licensed, but not yet interlinked; missing, thus, opportunities for data analytics. Only dedicated efforts like openlaws[3] are fully exploiting this potential. The LKG will comprehensively integrate these resources assuming the linked data paradigm, containing both structured and unstructured content, the latter being annotated with terminologies and links to other documents. Most legal documents at European (e.g., legislation in Eur-Lex) and national levels (e.g. in CENDOJ[4], Spanish case law from the Supreme Court, National High Court and other organs) already have some unstructured text (e.g., the text of a directive) and structured metadata (e.g., document's authorship, date of creation).

Step 2. Integration: In the second step, the basic LKG management infrastructure will be implemented, as well as the set of common services which will include (i) linking and extraction services, for the interlinking and annotation of documents and data; (ii) language services, for the domain specific machine translation and summarisation of documents, and (iii) information retrieval and smart services, for the intelligent search, curation, and comparison of documents, as well as a service of customizable notifications and alerts.

Step 3: Exploitation: Three pilot B2B applications will be developed on top of the LKG and the common services to demonstrate the technical feasibility and business potential of the Lynx platform. In each of these pilots, customised solutions will be created to meet the specific needs of three business cases. These pilots are: (i) legal compliance solution for data protection, where data protection related documents are innovatively managed, analysed, and visualised across different jurisdictions; (ii) compliance assurance services in the Oil & Gas and Energy sectors, where the Lynx platform helps understanding regulatory regimes (norms and standards) related to operations; (iii) compliance solution for strategy design in labour law, where legal provision, case law, administrative resolutions, and expert literature are interlinked, analysed, and compared to define the strategy that is applicable or of interest for the case.

4. Expected Outcomes

The outcomes of this project will be:

Business outcomes: The companies involved in the project will be able to support their customers more effectively in the compliance-related services they provide, considering faster, more complete information, and reducing risks. This novel approach will transform the way these companies operate. The feasibility of using open-data based services for compliance will be demonstrated and other law firms and compliance assurance companies will be encouraged to adopt the Lynx approach, thus multiplying the impact and reducing the language and legal barriers in the fragmented EU markets and beyond.

Societal outcomes: EU citizens will have access to information related to compliance for the first time integrating legislation and standards in a single multilingual portal. Citizens moving or operating across borders will enjoy more opportunities because legal uncertainty and risks will be reduced.

Technical and innovation outcomes: New legal linked data, publicly offered through APIs and highly connected to external datasets. A set of core services that can be reused by third parties external to the project and three running pilots covering in depth specific domains, with a user interface designed to address specific business cases. The new point to access open legal information, different in business models to the traditional legal information providers, and a proof that open data properly curated and connected suffices in many of the standard needs of lawyers in the context of compliance SMEs will benefit from Lynx in two different flavours: companies registered in the Lynx platform and companies in the portfolio of customers of law firms and consultancy agencies registered in Lynx.

5. The Lynx platform and Business Cases

The Lynx platform is an information system based on semantic and multilingual data technologies and aimed at assisting companies in internationalisation processes through a set of compliance-oriented digital services.

As can be seen in figure 2, the platform will consists of a knowledge graph with legal and regulatory data named Legal Knowledge Graph, (block 1, at the bottom of the figure), a collection of common services for compliance (block 2, in the middle) and the user interfaces for three specific-domain pilots (block 3, at the top). The Lynx platform will also offered as a web portal accessible to the general public who can search, browse, and access open legal and regulatory documents.

A number of data source dependant converters will be used to ingest the documents and data in the Lynx platform. Such data will be very diverse in nature, scope and formats, and it will comprise: legal resources (legislation and case law), standards, language resources both domain-specific and domain-neutral (terminologies, dictionaries, vocabularies, etc.), external open data sources (such as

[1] https://www.europeandataportal.eu/
[2] http://eur-lex.europa.eu/homepage.html
[3] https://openlaws.com/
[4] http://www.poderjudicial.es

those contained in the cloud of linked open data), and company private documents.

The core of the Lynx platform is a set of common services that are built upon existing and well tested technologies developed by the different technological partners: (i) the PoolParty semantic middleware suite; (ii) the FREME framework for multilingual and semantic enrichment of digital content; (iii) DKT API for digital curation processes; (iv) TILDE custom machine translation and cloud terminology services APIs; (v) OEG-UPM' APIs for ontology engineering and linked data publication, and (vi) K Dictionaries dictionary APIs.

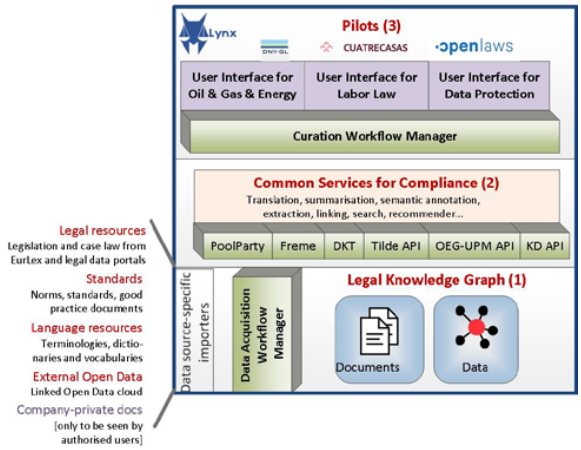


Figure 2. High level architecture of the Lynx platform

The common services provide core functionalities for annotating, linking, translating, and processing documents and data. Some of these services are invoked in order to treat new content arriving in the LKG (e.g., documents are annotated, linked or translated), and some of these services are used by the pilots, which are web-based applications to solve specific problems in three different domains. The invocation of these services for ingesting new content is coordinated by the Data Acquisition Workflow Manager, whereas their invocation for providing functionalities to the pilots is orchestrated by the Curation Workflow Manager. The common services are the backbone of a modular architecture which allows for the growing of input data sources and applications, as well as the dynamic scaling of the services in order to accommodate both surges and future expansion.

The platform will be instantiated in three pilots for each of the business cases considered in the project. The pilots will exploit different parts of the LKG and will make use of different common services according to their necessities. Each pilot will develop a specialised user interface that will capture the necessities of every business case according to their respective requirements.

These pilots support the vision of Lynx: that companies belonging to every sector of activity face compliance challenges when crossing the borders, and that an integrated approach to law and standards across different languages would increase competitiveness for companies in Europe reducing costs and corporate risks.

First, the Data Protection pilot will evidence the benefits of connecting legislation and case law from the EU and Members States. Second, the Oil & Gas and Energy pilot will demonstrate how the management of norms and industry standards is simplified by aggregating, comparing, and harmonising heterogeneous sources. Third, the Labour Law pilot will validate the appropriateness of the solution across different jurisdictions and languages, especially in such a complex domain as the Labour Law one, where each member state has different regimes, procedures, and standards.

6. Lynx Consortium

Ten partners from 7 different countries and complementary skills will work together to attain the objectives described in the previous sections. The Ontology Engineering Group from the Universidad Politécnica de Madrid is coordinating the project and is leading the Data Acquisition and Management work package. It is also contributing to the development of services, given its expertise in semantic technologies and data-driven language technologies.

The other academic partner, the Autonomous University of Barcelona, represented by the Institute of Law and Technology, will bring in its expertise in application of new technologies to the Legal Domain, and will lead the Industry requirements elicitation process, and the Dissemination and exploitation of project results.

The German Research Centre for Artificial Intelligence (DFKI), as one of the leading institutions in Europe for advanced IT applications dealing with human language, will lead the development of a set of curation tools, technologies and services to bridge between the core platform services and the use case specific pilots.

Semantic Web Company is an Austrian SME offering ICT consulting services and solutions in the fields of semantic information and data management. This company will lead the development of the Lynx platform core services, bringing in at the same time proprietary software components and semantic tools.

The specification of the technical architecture as well as the integration of all platform components will be performed by Alpenite, an IT software consulting and system integration company with headquarters in Italy.

The Latvian SME Tilde will bring in its expertise in multilingual natural language data processing technologies. Specifically, it will provide custom machine translation services and cloud terminology services, and will be also involved in other technological tasks, management, dissemination, and exploitation tasks.

The main provider of lexical data will be KDictionaries, an Israeli technology-driven-content creator that has developed quality lexicographic data for 50 languages.

Openlaws, another Austrian SME operating in the legal tech domain, will represent the use case on data protection, and will lead the development and integration of the two other pilots defined in the project. It will also contribute with core technology to the data acquisition and management tasks.

The use case on industry standards is led by DNV.GL, a standards certifying company with headquarters in Netherlands, Norway and Germany. This company will contribute to the requirements elicitation and specification for its pilot, and to the creation of the regulatory graph within the LKG.

Finally, the labour law pilot will be led by Cuatrecasas, a Spanish law firm with presence in over ten countries. It will also contribute with functional specification requirements and the development of the pilot.

9. Acknowledgements

Lynx has received funding from the Horizon 2020 European Union (EU) Research and Innovation programme under Grant Agreement: 780602. We thank all partners of the Lynx consortium for their contribution to the project's inception and proposal document. Special thanks to Ilan Kernerman and Georg Rehm for their suggestions to improve the paper.

7. Bibliographical References

Silveira, P., Rodríguez, C., Casati, C., Daniel, F., D'Andrea, V., Worledge, C. and Taheri, Z. On the design of compliance governance dashboards for effective compliance and audit management. In Service-Oriented Computing. ICSOC 2009, pp.208-217. Springer Berlin Heidelberg, 2010.

Curation Technologies for the Construction and Utilisation of Legal Knowledge Graphs

Julian Moreno Schneider, Georg Rehm
Language Technology Lab, DFKI GmbH
Alt-Moabit 91c, 10559 Berlin, Germany
julian.moreno_schneider@dfki.de, georg.rehm@dfki.de

Abstract

Under the umbrella of the European project LYNX we currently develop technologies for the construction of a legal knowledge graph and a corresponding system that makes use of this legal knowledge graph. The final platform will eventually bundle a set of semantic services into a curation technology system, which is meant to assist users to process legal and regulatory content and data more efficiently and more effectively. In this paper we present an overview of the current state of the art with regard to semantic technologies and natural language processing approaches applied to the legal domain.

Keywords: Curation Technologies, Natural Language Processing, Legal Domain

1. Introduction

The ever growing amount of digital information not only offers immense opportunities but also makes it necessary, in practically all professional areas and also niches, to develop new, efficient and effective approaches for processing digital content in order to make the information available in a way that fits the users' specific use cases as adequately as possible (Rehm et al., 2018). In a general sense, these professional users are the *curators* of digital content, for example, a journalist, producer of a television programme, a knowledge worker, a scholar, or someone who is collecting information to put together a report. The processes involved in digital curation include, among others, sorting, analysing, summarising, translating and paraphrasing digital content in terms of large amounts of incoming data and producing some kind of output, e. g., a study that relies on facts and figures found in a large data collection.

This data collection is typically a combination of publicly available sources (e. g., Wikipedia and other websites) and in-house collections owned by the respective organisation or data sets that the organisation has access to. The amount of time digital curators have so that they can familiarise themselves with new topics depends on the respective sector and is typically not a lot, ranging from a couple of hours or days to a few weeks at most. Working under intense time pressue, digital curators may not be able to identify and locate all relevant information contained in a sizable document collection (Neudecker and Rehm, 2016).

Ideally, digital curators should be able to explore, handle, analyse, summarise, translate, curate their data collections as quickly and efficiently as possible, enabling them to concentrate on producing the required output document or piece of information (Schneider et al., 2016; Bourgonje et al., 2016; Srivastava et al., 2016). The brief description given above is what we perceive as the core of any content curation system. In the highly dynamic legal domain we face the additional challenge of new decisions and dynamic updates where the law, or its interpretation, can change significantly with every court case.

Legal documents have a certain set of key characteristics (van Opijnen and Santos, 2017): high volume, extensive document length, very specific (internal) structure, heterogeneity of types, self-contained documents, hierarchy, temporal aspects, legal terminology, multilingualism, multi-jurisdictionality and, crucially, importance and abundance of citations and cross-references. All of these features make documents from the legal domain highly interesting and also challenging objects for a digital curation system. The idea is to analyse the documents automatically in order to provide added value based, among others, on semantically enriched documents – an important prerequisite for providing suitable curation services in different use cases. This is one of the objectives of the project LYNX ("Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe"), a three-year project, funded by the European Union, that consists of a consortium of ten partners.[1] Lynx aims to create a knowledge graph of legal and regulatory data towards compliance, in which heterogeneous data sources from different jurisdictions, languages and orders are aggregated and interlinked by a collection of advanced analysis and curation services.

In this article we provide an overview of the current state-of-the-art in curation technologies in the legal domain, concentrating on the following three questions:

1. What kind of (semantic) technologies are currently being used in production systems and research prototypes in the field of smart digital services for legal data, legal documents, etc.?
2. What kind of features and functionalities are currently explored in research labs and what is actually being used in terms of novel technologies?
3. What are the most important open research questions in Natural Language Processing and Language Technology for the legal domain, NLP for legal documents, processing legal information, automatically understanding and machine-reading the law, etc.?

The main contribution of this paper is a detailed description of previous research efforts and commercial tools, such as

[1] http://www.lynx-project.eu

curation systems and technologies, the application of curation technologies to a new, concrete and specific domain, i. e., legal information systems. The remainder of this article is structured as follows. Section 2. provides a summary of curation systems in the legal domain, both commercial and prototypical. In Section 3., an overview of important research areas in the legal domain with regard to Natural Language Processing (NLP, Section 3.1.) and semantic technologies (Section 3.2.) is presented. Section 4. concludes the article.

2. Curation Systems in the Legal Domain

Even if they do not use this specific name, curation technologies have been in use in several different domains, including the legal area. Here, the uptake has been a bit slower than in other domains because, in many countries, collections of legal documents and data sets have been monopolised by commercial enterprises which means that there are access restrictions on multiple levels.

2.1. Commercial Systems and Services

One of the most visible companies in the area of semantic technologies and services for the legal domain is LexisNexis.[2] Their system is the market leader and offers services for the legal domain, such as legal research, practical guidance, company research and media-monitoring solutions, intellectual property, litigation strategy and discovery, practice and legal department management as well as compliance and due diligence among others.

Also visible in the legal area is WestLaw, an online service that allows legal professionals to find and consult the needed legal information.[3] Developed by Thomson Reuters, one of the goals of Westlaw is to enable professionals to put together the strongest argument possible.

Apart from these two providers, there are other smaller companies and services that offer legal research solutions and analytic environments, such as RavelLax,[4] which "provides services designed to help legal professionals draw insights and connections using advanced analytical algorithms", or Lereto[5], offering tools for legal document processing. A commercial search engine for legal documents, iSearch, is a service offered by LegitQuest.[6]

The Casetext CARA Research Suite allows uploading a brief and then retrieving, based on its contents, useful case law.[7] In its own words, CARA is an AI-backed automated research assistant, empowering litigators to better serve their clients through advanced information services, supported through technology, and expert analysis from the legal community. CARA's contextual search to help litigators get the answers they need fast so they can spend more time on higher-value work. The company is comprised of lawyers, data scientists, engineers, and designers helping attorneys better to represent their clients.

In addition to these products and services there is also a growing number of startup companies active in the legal domain – from applying AI techniques for automatically analysing large amounts of documents to better supporting communication among law firms, clients and stakeholders.

2.2. Research Prototypes

While there are several research prototypes that can be considered "curation systems for the legal domain", most of the documented systems were developed in the 1990s under the umbrella of Computer Assisted Legal Research – CALR (Span, 1994). In the following we briefly review several of these systems.

Most prototypes we have been able to find in the literature are not curation systems per se (meaning, in the sense described above) but systems that offer a very specific functionality that a legal document curation system (in our sense) would offer together with many other functionalities. One example is the open source software for the analysis and visualisation of networks of Dutch case law presented by (van Kuppevelt and van Dijck, 2017). This technology assists in answering legal research questions by means of determining relevant precedents (analysing the citation network of case law), comparing them with those identified in the literature, and determining clusters of related cases. Another prototype extracting references from legal documents is described by (Agnoloni et al., 2017). They introduce a framework for the extraction of legal references from case-law of European Member States based on an approach applicable to multiple languages and jurisdictions, helping national data providers to reduce the effort needed to develop their own extraction solution. (Gifford, 2017) propose a search engine for legal documents where arguments are extracted from appellate cases and are accessible either through selecting nodes in a litigation issue ontology or through relational keyword search.

A relevant curation prototype is Lucem (Bhullar et al., 2016), a web-based system that provides a solution for obtaining legal information in an accessible and intuitive way. The system tries to mirror the way lawyers approach legal research, developing visualisations that provide lawyers with an additional tool to approach their research results. Eunomos is a curation prototype that uses NLP techniques to semi-automate the construction and analysis of knowledge. This legal knowledge management service enables users to view legislation from various sources and to find the relevant definitions and explanations of legal concepts in a given context (Boella et al., 2012). Functionalities included are the ability to view legislation at European, national and regional level, links between different parts of legislation, lists of similar legislation, a mechanism for classifying norms in user-defined categories and a notification service that alerts users when a newly downloaded legislation appears.

3. Important Research Areas

To identify current research strands and trends we checked the scientific programme of the most relevant conferences in the area to identify common topics. The conferences in

[2] https://www.lexisnexis.com

[3] http://legalsolutions.thomsonreuters.com/law-products/westlaw-legal-research/

[4] http://ravellaw.com

[5] https://www.lereto.at

[6] https://www.legitquest.com

[7] https://casetext.com

question are AI4J[8], JURIX[9], JURISIN[10] and ICAIL (International Conference on AI and Law)[11].
There are several overarching topics that are recurrent among the conferences listed above. These are reasoning and inference, argumentation extraction, evidential reasoning, legal interpretation, decision making, extraction of structure and connections of legal texts and rules, annotation, information retrieval and discovery, text classification, summarisation, translation, linked data and open data, knowledge acquisition, natural language processing, legal knowledge representation, including legal ontologies and common sense knowledge.

3.1. Natural Language Processing for the Legal Domain

Within the broad field of NLP, research currently focuses upon the topics briefly reviewed below.

3.1.1. Citation Analysis

Almost all types of documents that belong to the legal domain refer to laws, paragraphs, rules, correspondence or arbitrary other documents, which is why citation and cross-reference analysis is an almost mandatory step in any processing pipeline. There is a multitude of approaches focused on citation analysis, addressing the challenge from different perspectives and with different methods. There appear to be two major directions, i. e., applying network analysis to citations (Zhang and Koppaka, 2007), (Winkels et al., 2011), (Lupu and Voeten, 2012), (Neale, 2013) and classification systems estimating the status of the cited case (Galgani et al., 2015). (Zhang and Koppaka, 2007) develop a semantics-based legal citation network, which is a tool that extracts and summarises citation information into a network, allowing the users to navigate the citation network and to study how citations are interrelated and how legal issues have evolved in the past. LEXA (Galgani et al., 2015) is a system that relies on Ripple Down rules approach to identify citations within the "distinguished" class. This category is generally best linguistically signaled and is therefore suitable for achieving high precision and recall.

3.1.2. Argument Extraction and Mining

Like citation analysis, argument extraction is an important part in the understanding of legal documents. Recognising the arguments used in case law is vital for identifying similar arguments in other documents and to predict possible outcomes of a specific case. Many different approaches have been applied, such as statistical methods over annotated corpora, used by (Moens et al., 2007) to automatically detect sentences that are part of a legal argument. (Cabrio et al., 2016) summarise current trends in argumentation mining and discuss future challenges.

3.1.3. Reasoning

Logical reasoning is, naturally, an important part of a legal expert's day-to-day work, which is why there have been several attempts at performing automatic reasoning techniques based on a specific set of information or knowledge provided. As stated by (Vlek et al., 2014), there are three main approaches to performing reasoning with or over evidence: argumentative, narrative and probabilistic approaches. (Vlek et al., 2014) combine these approaches to form a design method for constructing a Bayesian network based on narratives. An extension of this work, (Vlek et al., 2016), proposes a method combining a probabilistic approach with a narrative approach to reasoning with legal evidence. Whereas a Bayesian network is a popular tool for analysing parts of a case, the advantage of a narrative approach is that it provides the global perspective on the case as a whole. (Verheij, 2017) use a different approach, in which they propose a formalism, in which the validity of arguments is defined in terms of case models.

3.1.4. Summarisation

Many researchers emphasise that the average length of documents in the legal domain is rather extensive, plus, one case usually comprises many different documents of several different types (van Opijnen and Santos, 2017). This is why it is, for legal experts, a difficult challenge to acquire first an overview and then detailed knowledge of the content of all of these documents. Automatic summarisation could help lawyers to familiarise themselves quickly and efficiently with a new set of documents on a specific case.
A common approach in automatic summarisation, also used in the legal domain, is sentence classification and sentence ranking. The SUM project (Grover et al., 2003) applied automatic summarisation to the legal domain by means of sentence classification based on the sentences' rhetorical roles. They explored the relationship between linguistic features and argumentative roles in order to classify sentences. Another approach using sentence classification is the prototypical summarisation system, LetSum (Legal text Summarizer) (Farzindar and Lapalme, 2004). It classifies sentences into four themes: introduction, context, juridical analysis and conclusion. Summaries are generated in four steps: thematic segmentation, filtering to eliminate unimportant quotations and noise, selection of candidate units and generation of the summary.
(Polsley et al., 2016) use a sentence classification method, that is based on word frequency augmented with domain-specific knowledge. They implemented a tool called CaseSummarizer, whose processing pipeline consists of three steps: preprocessing, scoring of sentence relevance, and domain processing. They present summaries to the user through a multi-faceted interface with abbreviations, significance heat maps, and other flexible controls.
(Yousfi-Monod et al., 2010) use supervised machine learning for summarising legal documents based on a Naive Bayes classifier. They use a set of surface, emphasis, and content features. For the training of these machine learning based approaches, annotated data is needed. For data acquisition, a corpus of UK House of Lords judgments[12] is created (Grover et al., 2004). It contains three layers: rhetorical status annotation, detailed linguistic markup, and relevance annotation.

[8] http://www.ai.rug.nl/~verheij/AI4J/
[9] https://jurix2017.gforge.uni.lu
[10] http://research.nii.ac.jp/~ksatoh/jurisin2017/
[11] https://nms.kcl.ac.uk/icail2017/
[12] http://www.ltg.ed.ac.uk/SUM/

3.1.5. Information Retrieval

Given the large amount of information handled in legal cases, it is essential to have good search and retrieval capabilities. Many researchers focus on improving search engines in this domain. Two approaches to legal IR, based on manual knowledge engineering (KE) and NLP, are presented and compared in (Schafer and Maxwell, 2008). They concluded that IR based solely on KE is not sustainable in the long run.

The ontology-based IR system EgoIR is presented by (Gómez-Pérez et al., 2006). It aims to retrieve government documents in a timely and accurate manner. Ontologies are used for two purposes: to guide users to the legal terms, enabling them to avoid mistakes at constructing a query and to improve interoperability in legal applications.

Apart from the topics mentioned above, there are many additional questions being investigated. A sentence classification approach in the legal domain is presented by (van Opijnen and Santos, 2017; Shulayeva et al., 2017), where a set of linguistic features (part of speech tags, unigrams, dependency pairs, length of the sentence, position in the text and cita, which indicates whether there is a citation instance in the sentence) is extracted using NLTK (Loper and Bird, 2002) and CoreNLP (Manning et al., 2014), later on to classify the sentence with WEKA (Hall et al., 2009).

3.2. Semantic Technologies

The legal domain is characterised by having an incredibly large number of established terms. There have been several attempts to organise these terms in ontologies and semantic systems, which is why there is a lot of research related to semantic technologies including ontology bootstrapping and generation, ontology population and the use of ontologies for IR and semantic annotation.

Some common approaches for the population of ontologies use standard NLP tools (such as TreeTagger, GATE, YaTeA, etc.) or ontology learning tools (Lehmann and Voelker, 2014) (such as OntoGen, ASIUM, Text-To-Onto, Text2Onto and TERMINAE). (El Ghosh et al., 2017) use the methodology Terminae (Aussenac-Gilles et al., 2000) for legal ontology population based on two approaches: top-down and botton-up. The bottom-up approach uses linguistic information (using YaTeA) for extracting features (concepts and relations) and to convert them into domain-specific ontologies. The top-down approach is based on the definition and (partial) reuse of existent ontologies.

(Francesconi et al., 2010) perform legal knowledge acquisition based on top-down and bottom-up approaches. They present a methodology for multilingual legal knowledge acquisition and modeling. The top-down approach is the definition of the conceptual structure of the legal domain on the basis of expert judgments. This structure is language-independent, modeled as an ontology, and can be aligned with other ontologies that capture similar or complementary knowledge, in order to provide a wider conceptual embedding. The bottom-up approach is a linguistic text-based population of conceptual structures using semi-automatic NLP techniques, which maximise the completeness and domain-specificity of the resulting knowledge. A different approach using semantic information is SALEM, the automatic enrichment of legal texts with semantic annotations (Biagioli et al., 2005). SALEM is an NLP system for the classification and semantic enrichment of articles of law. The enrichment helps effectively index and retrieve legal documents. It classifies paragraphs according to their regulatory content and extracts relevant text fragments corresponding to specific semantic roles. The ontology distinguishes three categories: obligations, definitions and modifications.

In addition, knowledge representation is an important topic in the legal domain. (Perinan-Pascual and Arcas-TÃ, 2014) define a knowledge representation model inside the frame of FunGramKB, a lexical-conceptual multilingual knowledge graph. The generation of the knowledge graph is divided into five steps: (1) definition of filters, (2) corpus indexing, (3) n-gram and statistics extraction, (4) terms identification and (5) corpus validation. A proof of concept for the ontological representation of normative requirements as Linked Data on the Web is proposed by (Gandon et al., 2017), who present an extension of the LegalRuleML ontology to model normative requirements and rules.

Furthermore, there are multiple available ontologies for the legal domain. (Breuker et al., 2009) provide a corresponding list. Several examples are FOLaw – Functional Ontology of Law (Valente et al., 1994), OPLK – Ontology of the Professional Legal Knowledge (Benjamins et al., 2004), Jur-Wordnet (Gangemi et al., 2003), DALOS (Francesconi and Tiscornia, 2008) and OPJK – Ontology of Professional Judicial Knowledge (Casanovas et al., 2009).

An ontology learning system (T2K) that includes NLP tools, statistical text analysis and machine learning is used by (Lenci et al., 2007). Their approach allows the dynamic integration of new modules to provide an incremental representation of the content of vast repositories of unstructured documents. They also include bootstrapping techniques to develop more sophisticated levels of content representation starting from knowledge-poor language tools.

(Casanovas et al., 2016) provide an overview of a special issue of the journal Semantic Web, aimed at the legal domain, summarising research carried out in the legal domain in the last 15 years. They emphasise five ontology definition and generation approaches: (1) an OWL ontology making it possible to describe a judge's interpretations of the law while engaging in the legal reasoning on which basis a case is adjudicated (Ceci and Gangemi, 2016). (2) an OWL ontology, framed in CELLAR, for describing normative provisions to enable advanced access to legal documents (Francesconi, 2016). (3) the LOTED2 ontology for the representation of European public procurement notices, enabling legal reasoning (Distinto et al., 2016). (4) the PPROC ontology, which enables the description of procurement processes and contracts (Muñoz-Soro et al., 2016). (5) the MPEG-21 Media Contract Ontology, which enables the description of contracts dealing with rights to multimedia assets and with any content protected by intellectual property (Rodríguez-Doncel et al., 2016).

4. Conclusions

This article presents an overview of approaches that are highly relevant for the development of a system for the

curation of content and documents from the legal domain, aimed at the construction and utilisation of legal knowledge graphs. The article is structured into two parts. The first part presents existing curation systems, both research prototypes and commercial systems. The commercial market is currently dominated by two major players and several smaller companies (including several startups) that try to penetrate the market.

In our desk research we have found only very few non-commercial and/or free systems – prototypes are rarely used outside the laboratory. Among the main reasons for this situation is the fact that many important data and document collections are controlled by commercial companies and because of privacy and data protection issues. These issues are so severe that the situation is unlikely to change. Legal document collections contain very high numbers of names and events that many would not want and proabably also cannot be made public – the publication of a collection in the legal domain that has previously been anonymised has little or no value for the development of functional technologies. Despite a very high amount of research activity in the legal domain, this effort does not immediately translate into prototypes or free systems that are in widespread use.

The second part summarises current research strands in this area and analyses the main conferences on the topic. These research lines are divided into two main groups: NLP and semantic approaches. Regarding NLP, there are several interesting topics for the legal domain, such as reasoning, argument mining, summarisation and document linking. In the case of semantic approaches, the variety of topics is not as rich, and there are mainly three main topics: knowledge base construction, mainly based on existing ontologies, knowledge base population, mainly based on (semi-)automatic NLP and ontology learning, and semantic enrichment of documents.

This contribution has been prepared under the umbrella of the EU project LYNX, which has started in December 2017. The analysis of available technologies and current research strands will inform the design and development of a system, which makes use of curation technologies for the construction and utilisation of a legal knowledge graph.

5. Acknowledgements

The project LYNX has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 780602. More information is available online at http://www.lynx-project.eu.

6. Bibliographical References

Agnoloni, T., Bacci, L., Peruginelli, G., van Opijnen, M., van den Oever, J., Palmirani, M., Cervone, L., Bujor, O., Lecuona, A. A., García, A. B., Caro, L. D., and Siragusa, G. (2017). Linking european case law: BO-ECLI parser, an open framework for the automatic extraction of legal links. In Wyner and Casini (Wyner and Casini, 2017), pages 113–118.

Aussenac-Gilles, N., Biébow, B., and Szulman, S. (2000). Corpus analysis for conceptual modelling.

Benjamins, V. R., Contreras, J., Casanovas, P., Ayuso, M., Becue, M., Lemus, L., and Urios, C. (2004). Ontologies of professional legal knowledge as the basis for intelligent it support for judges. *Artificial Intelligence and Law*, 12(4):359–378.

Bhullar, J., Lam, N., Pham, K., Prabhakaran, A., and Santillano, A. J., (2016). *Lucem: A Legal Research Tool*. Number 63. Computer Engineering Senior Theses.

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., and Soria, C. (2005). Automatic semantics extraction in law documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, pages 133–140, New York, NY, USA. ACM.

Boella, G., di Caro, L., Humphreys, L., Robaldo, L., and van der Torre, L. (2012). Nlp challenges for eunomos, a tool to build and manage legal knowledge.

Bourgonje, P., Schneider, J. M., Rehm, G., and Sasaki, F. (2016). Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In Aldo Gangemi et al., editors, *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 13–16, Edinburgh, UK, September. The Association for Computational Linguistics.

Breuker, J., Casanovas, P., Klein, M. C. A., and Francesconi, E. (2009). *Law, ontologies and the Semantic Web: Channelling the legal information flood*. IOS Press, Amsterdam.

Cabrio, E., Hirst, G., Villata, S., and Wyner, A. (2016). Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (Dagstuhl Seminar 16161). *Dagstuhl Reports*, 6(4):80–109.

Casanovas, P., Casellas, N., and Vallbé, J.-J. (2009). An ontology-based decision support system for judges. In *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 165–175. IOS Press.

Casanovas, P., Palmirani, M., Peroni, S., van Engers, T. M., and Vitali, F. (2016). Semantic web for the legal domain: The next step. *Semantic Web*, 7(3):213–227.

Ceci, M. and Gangemi, A. (2016). An owl ontology library representing judicial interpretations. *Semantic Web*, 7(3):229–253.

Distinto, I., d'Aquin, M., and Motta, E. (2016). Loted2: An ontology of european public procurement notices. *Semantic Web*, 7(3):267–293.

El Ghosh, M., Naja, H., Abdulrab, H., and Khalil, M. (2017). Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules. *Procedia Computer Science*, 112:632–642.

Farzindar, A. and Lapalme, G., (2004). *LetSum, an automatic Legal Text Summarizing system*, pages 11–18. IOS Press, Berlin, dec.

Francesconi, E. and Tiscornia, D. (2008). Building semantic resources for legislative drafting: The dalos project.

Francesconi, E., Montemagni, S., Peters, W., and Tiscornia, D., (2010). *Integrating a Bottom–Up and Top–Down Methodology for Building Semantic Resources for*

the Multilingual Legal Domain, pages 95–121. Springer Berlin Heidelberg, Berlin, Heidelberg.

Francesconi, E. (2016). Semantic model for legal resources: annotation and reasoning over normative provisions. *Semantic Web*, 7(3):255–265.

Galgani, F., Compton, P., and Hoffmann, A. (2015). Lexa: Building knowledge bases for automatic legal citation classification. *Expert Systems with Applications*, 42(17):6391–6407.

Gandon, F., Governatori, G., and Villata, S. (2017). Normative requirements as linked data. In Wyner and Casini (Wyner and Casini, 2017), pages 1–10.

Gangemi, A., Sagri, M.-T., and Tiscornia, D. (2003). Metadata for content description in legal information. In *Procs. of LegOnt Workshop on Legal Ontologies*.

Gifford, M. (2017). Lexridelaw: an argument based legal search engine. In *ICAIL '17*.

Gómez-Pérez, A., Ortiz-Rodriguez, F., and Villazón-Terrazas, B. (2006). Ontology-based legal information retrieval to improve the information access in e-government. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 1007–1008, New York, NY, USA. ACM.

Grover, C., Hachey, B., Hughson, I., and Korycinski, C. (2003). Automatic summarisation of legal documents. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, ICAIL '03, pages 243–251, New York, NY, USA. ACM.

Grover, C., Hachey, B., Hughson, I., and Place, B. (2004). The holj corpus: supporting summarisation of legal texts. In *In Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Lehmann, J. and Voelker, J. (2014). An introduction to ontology learning. *Perspectives on Ontology Learning. IOS Press, Amsterdam, The Netherlands*.

Lenci, A., Montemagni, S., Pirrelli, V., and Venturi, G. (2007). Nlp-based ontology learning from legal texts. a case study. In Pompeu Casanovas, et al., editors, *LOAIT*, volume 321 of *CEUR Workshop Proceedings*, pages 113–129. CEUR-WS.org.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lupu, Y. and Voeten, E. (2012). Precedent in international courts: A network analysis of case citations by the european court of human rights. *British Journal of Political Science*, 42(2):413–439.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM.

Muñoz-Soro, J. F., Esteban, G., Corcho, O., and Serón, F. (2016). Pproc, an ontology for transparency in public procurement. *Semantic Web*, 7(3):295–309.

Neale, T. (2013). Citation analysis of canadian case law. *J. Open Access L.*, 1:1.

Neudecker, C. and Rehm, G. (2016). Digitale Kuratierungstechnologien für Bibliotheken. *Zeitschrift für Bibliothekskultur 027.7*, 4(2), November.

Perinan-Pascual, C. and Arcas-TÃ, F. (2014). La ingenierÃa del conocimiento en el dominio legal: La construcciÃ de una OntologÃa SatÃ©en FunGramKB. *Revista signos*, 47:113 – 139, 03.

Polsley, S., Jhunjhunwala, P., and Huang, R. (2016). Casesummarizer: A system for automated summarization of legal texts. In *COLING*.

Rehm, G., Schneider, J. M., Bourgonje, P., Srivastava, A., Fricke, R., Thomsen, J., He, J., Quantz, J., Berger, A., König, L., Räuchle, S., Gerth, J., and Wabnitz, D. (2018). Different Types of Automated and Semi-Automated Semantic Storytelling: Curation Technologies for Different Sectors. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 232–247, Cham, Switzerland, January. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Rodríguez-Doncel, V., Delgado, J., Llorente, S., Rodríguez, E., and Boch, L. (2016). Overview of the mpeg-21 media contract ontology. *Semantic Web*, 7(3):311–332.

Schafer, B. and Maxwell, T., (2008). *Concept and Context in Legal Information Retrieval*, pages 63–72. Frontiers in Artificial Intelligence and Applications. IOS Press.

Schneider, J. M., Bourgonje, P., Nehring, J., Rehm, G., Sasaki, F., and Srivastava, A. (2016). Towards Semantic Story Telling with Digital Curation Technologies. In Larry Birnbaum, et al., editors, *Proceedings of Natural Language Processing meets Journalism – IJCAI-16 Workshop (NLPMJ 2016)*, New York, July.

Shulayeva, O., Siddharthan, A., and Wyner, A. (2017). Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1):107–126, Mar.

Span, G. (1994). Lites: An intelligent tutoring system shell for legal education. *International Review of Law, Computers & Technology*, 8(1):103–113.

Srivastava, A., Sasaki, F., Bourgonje, P., Moreno-Schneider, J., Nehring, J., and Rehm, G. (2016). How to Configure Statistical Machine Translation with Linked Open Data Resources. In Joao Esteves-Ferreira, et al., editors, *Proceedings of Translating and the Computer 38 (TC38)*, pages 138–148, London, UK, November. Editions Tradulex.

Valente, A., Breuker, J., et al. (1994). Ontologies: The missing link between legal theory and ai & law. *Legal knowledge based systems JURIX*, 94:138–150.

van Kuppevelt, D. and van Dijck, G. (2017). Answering legal research questions about dutch case law with network analysis and visualization. In Wyner and Casini (Wyner and Casini, 2017), pages 95–100.

van Opijnen, M. and Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1):65–87, Mar.

Verheij, B. (2017). Proof with and without probabilities. *Artificial Intelligence and Law*, 25(1):127–154, Mar.

Vlek, C. S., Prakken, H., Renooij, S., and Verheij, B. (2014). Building bayesian networks for legal evidence with narratives: a case study evaluation. *Artificial Intelligence and Law*, 22(4):375–421, Dec.

Vlek, C. S., Prakken, H., Renooij, S., and Verheij, B. (2016). A method for explaining bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, 24(3):285–324, Sep.

Winkels, R., Ruyter, J. d., and Kroese, H. (2011). Determining authority of dutch case law.

Adam Z. Wyner et al., editors. (2017). *Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017*, volume 302 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Yousfi-Monod, M., Farzindar, A., and Lapalme, G. (2010). Supervised machine learning for summarizing legal documents. In Atefeh Farzindar et al., editors, *Advances in Artificial Intelligence*, pages 51–62, Berlin, Heidelberg. Springer Berlin Heidelberg.

Zhang, P. and Koppaka, L. (2007). Semantics-based legal citation network. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 123–130, New York, NY, USA. ACM.

Towards a Workflow Manager for Curation Technologies in the Legal Domain

Julian Moreno Schneider, Georg Rehm
Language Technology Lab, DFKI GmbH
Alt-Moabit 91c, 10559 Berlin, Germany
julian.moreno_schneider@dfki.de, georg.rehm@dfki.de

Abstract

We develop a system for the curation and further processing of documents from the legal domain. The platform is based on a legal knowledge graph. The overall project will result in three use-case-specific prototypes for different areas of the legal domain. For the purpose of designing the exact needs, demands, ideas, wishes and feature requests we currently collect the functional and non-functional requirements from the three use case partners. The objective of our work is the design and implementation of a generic, yet customisable, workflow management system for content and data curation services in the legal domain. In this article we describe and discuss how the inherent characteristics of a specific domain influence the design and development process of automatic workflows of text and data processing as well as curation components. Different techniques for the analysis and for collecting requirements are presented, followed by our survey and hybrid approach.

Keywords: curation technologies, requirements gathering, workflow definition

1. Introduction

European enterprises that operate internationally, especially small and medium-sized companies (SMEs), face multiple difficulties when attempting to offer and to market their products and services in other countries. Complying with regulatory and legal aspects is a hard challenge, which is usually delegated to law firms and consultancies. They have to identify, retrieve and process documents in multiple languages, from various sources and published by various institutions according to different criteria and formats. The further expansion and internationalisation of European SMEs is severely hindered by this situation. The potential of smart technologies to address the situation and to support these companies is enormous.

Current content and data analysis solutions are mature enough to be transferred to the market and to benefit from the new opportunities created by the Linked Data paradigm and the Open Data movement. Among these mature solutions are curation technologies, that enable and support the semantic analysis of documents with the help of automatic processes (Bourgonje et al., 2017) in order to extract information and to enrich single documents and whole document collections (Bourgonje et al., 2016). The goal is to make knowledge workers, who process and make use of these documents, more efficient and more effective in their day to day work, supporting them by delegating tasks that can be automised to the machine (summarisation, translation, report generation, named entity recognition, time expression analysis etc.) (Rehm et al., 2017a; Schneider et al., 2017; Rehm et al., 2017b; Rehm et al., 2018).

Documents that belong to the legal domain are highly interesting. Many types of legal documents exhibit rather fixed and clearly defined structures, which are typically considered an advantage when it comes to automatic analyses. Legal documents also contain multiple references to other documents, which make them difficult to read and fully comprehend. Often it is simply not feasible to read all documents referenced in a document. In addition to the high number of internal and external references, the ever-changing nature of law itself makes it important to have technologies that are capable of identifying these changes and reporting them whenever changes occur.

The objective of LYNX (Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe), a 36 months European Union project that started in December 2017, is the generation of a legal knowledge graph that contains different types of legal and regulatory data.[1] A set of advanced semantic services is currently under development to collect, to aggregate and to interlink data from heterogeneous sources and different jurisdictions, languages and orders. The project will eventually offer compliance-related functionalities that will be tested and validated in three use cases. The first pilot will be a legal compliance solution, where documents related to data protection are innovatively managed, analysed, and visualised across different jurisdictions. In the second pilot, LYNX will support the understanding of regulatory regimes, including norms and standards, related to energy operations. The third pilot will be a compliance solution in the domain of labour law, where legal provisions, case law, administrative resolutions, and expert literature will be interlinked, analysed, and compared to define legal strategies for legal practice.

In this article we describe the first steps towards the design and development of the underlying curation workflow manager by studying and analysing the requirements of the three pilots mentioned above. There are several important research questions that have to be answered to identify the needs of the three pilots:

1. Which are the specific needs of each use case?
2. Which datasets and common services are needed in each use case?
3. How can data and content be best organised and managed in the system so that the use case and corresponding pilot can be implemented?

[1] http://www.lynx-project.eu

4. Which services are needed? In which order and with which data and content will they be used?

5. What is the expected output of each use case?

In the EU project LYNX we apply curation technologies, applied to documents of several other domains in previous projects, to the legal domain. The main contribution of our work is a description of the first steps in the process of defining the workflows governing the curation processes, concretely, the requirements gathering process applied towards the definition of workflows in the legal domain.

The remainder of the article is structured as follows. Section 2. describes the concept of curation workflows and their application in the legal domain. Section 3. describes the solution for the requirements gathering and workflow definition processes. Section 4. concludes the article.

2. Curation Workflow Manager

A workflow is typically defined as "an orchestrated and repeatable pattern of business activity enabled by the systematic organization of resources into processes that transform materials, provide services, or process information" (BPM, 2009). This business activity is usually restricted to automatic processes. In our case we do not want to limit ourselves to automatic processes, but also include manual or intellectual activities carried out by experts. Therefore, the definition of workflows is not just a pipeline of automated components, but a complex structure or network of domain-specific steps.

For example, Figure 1 shows a workflow defined for the discovery and monitoring of legal documents. It is composed of seven tasks, of which four are automatic (blue boxes) and three are manual (orange boxes). The links between the tasks (their execution) depend on fulfilling the conditions established by the links themselves. Next to each task, a green box denotes the entities or roles involved in that task (human experts in case of manual tasks or systems in case of automatic ones).

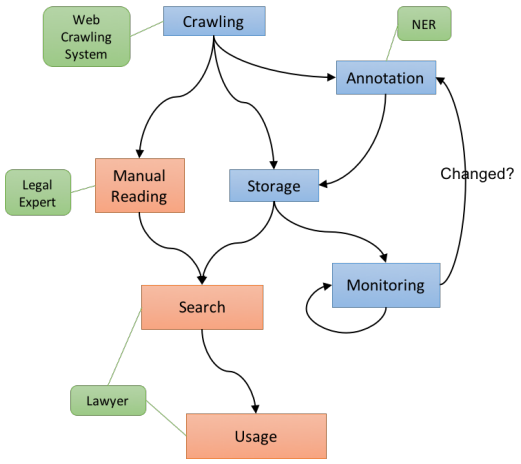


Figure 1: Workflow for the discovery and monitoring of legal documents inside an external database

This workflow is just an illustrative example, though, and we have only included the minimum tasks needed. A genuine workflow established in a company can be vastly more complex. We collaborate with three use case partner companies to get a better insight as to which are the specific workflows they currently use and follow and the requirements they have. Only with sufficient knowledge and insights into their processes and workflows can we begin the design of the curation workflow manager including its requirements and functionalities.

To provide a few other examples regarding the curation of digital documents and content in other domains, experts in a digital agency build mobile apps or websites for clients who provide the digital agency with documents, data, pictures, videos and other assets that are processed, sorted, augmented, arranged, packaged and then deployed. Journalists need to stay on top of the news stream including blogs, microblogs, newswires, websites etc. in order to produce a new article on a breaking topic, based on the information they collected, processed, sorted, evaluated, verified and synthesised (Rehm et al., 2017a).

The main tasks, apart from obtaining, analysing and organising documents, in data protection compliance are searching, browsing and commenting on documents. In regulatory compliance, companies have to support the current audit, verification and certification (including classification) of documents, apart from summarising larger documents and enabling commenting on them. Translating documents from other jurisdictions and comparing them among jurisdictions is also a common task. A labour law expert needs to access, aggregate and interlink relevant legal information, which starts by looking for relevant documents, manually creating links between legal provisions, case law, administrative resolutions and expert literature even across different jurisdictions, identifing relevant documents that may affect the case and tracing their changes, through the life cycle of the case.

The research and innovation project LYNX is currently in its first steps. Within LYNX, we design, define and develop a set of workflows for fulfilling the needed functionalities in the pilot use cases. The Curation Workflow Manager will be defined, including its main functionality to arrange, organise, orchestrate and combine all components in the platform in order to generate suitable workflows for every use case. The Curation Workflow Manager is located in between the pilot use cases and the common services.

Due to the fact that the platform will be based on a flexible service-oriented architecture, in which the basic services form a set of single services and data endpoints, the workflows act like a semantic middleware that integrates the core services in the use cases. This approach has the advantage of clearly separating the development of the distributed components of services from the actual use cases so that the different partners can develop their technologies at the same time without compromising the concurrent development work of other partners (as long as the agreed upon REST APIs remain stable and unchanged).

Workflows are sets of tasks (both manual and automatic) that are interconnected to complete a larger, more complex task. Among the automatic services that will become avail-

able during the course of the project and that will be used in the definition of the curation workflows are: named entity recognition, terminology identification, geolocation annotation, time expression analysis, translation, summarisation, semantic annotation, information extraction, linking and mapping, smart search, recommendation, alerts.
A comprehensive introduction to workflow management is offered by (Van Der Aalst and Van Hee, 2004), who present a basic overview of workflow terminology and organisation. The state of the art is presented by (Unertl, Kim M and Novak, Laurie L and Johnson, Kevin B and Lorenzi, Nancy M, 2010). Sequences of tasks and whole workflows are defined on a regular basis in many domains, which is why examples can be found in many different areas, among others, biomedical (BioNLP UIMA Component Repository (Baumgartner et al., 2008), JULIE Lab's UIMA Component Repository (JCoRe) (Hahn et al., 2008), Smntx (Chard et al., 2011), (Rak et al., 2012) or (Köster and Rahmann, 2012)), software development (Apache Oozie (Islam et al., 2012)) and in NLP, where many different frameworks for the definition of workflows exist: Taverna (Hull et al., 2006), Galaxy (Blankenberg et al., 2010), GATE (General Architecture for Text Engineering) (Cunningham et al., 2002), DKPro Core (de Castilho and Gurevych, 2014), U-Compare (Kano et al., 2009; Kano et al., 2011) and TextFlows (Perovšek et al., 2016).

3. Design, Definition, Development

The Curation Workflow Manager (CWM) is a component of the architecture that is responsible for orchestrating and managing the workflows and adapting them to the specific needs of the three use cases. In this project, the users of the system are the actual use case partners including, if applicable, their clients and other immediate stakeholders. Therefore, we need to collect and specify the workflows that are currently used and that are to be used in the future, given the new LYNX functionalities, and how we can realise them using the above mentioned semantic services.
The first step towards the design and definition of the CWM is a list of requirements obtained from the users. To that end, many different techniques can be applied. Some examples are: Document Analysis (evaluating the documentation of a present system), Feasibility Study (studying existing systems and the possibility of replacing them), Interview (with one more future users), Observation (studying users in their workplace), Prototyping (gathering preliminary requirements to build an initial prototype), Survey/Questionnaires (gathering information from a small or large amount of users), Brainstorming (identifying all possible solutions to problems) and Requirements Workshop (more organised and structured than a brainstorming session). A complete description of requirement gathering techniques can be found in (Fricker et al., 2015).
Based on the existing and commonly used techniques, and also taking the typical constraints of a research project into account, we opt for a hybrid approach that consists of the following steps:

- First we define a survey the main objective of which is to collect a first set of requirements, needs, ideas and visions the use case partners have.
- Second, we use the results obtained from the survey to design a first, still coarse-grained specifications of the workflows for each of the three use cases.
- Based on the three sets of coarse-grained specifications we plan several brainstorming workshops, in which we will collect the requirements on a much more detailed level from the partners. There are several options how to organise these workshops. One fundamental distinction relates to the question if a new, even conceptually, GUI needs to be implemented or if the semantic services are to be integrated in existing systems and GUIs. In addition, if the survey results from two or maybe all three use case partners are similar, there may be no need for bilateral workshops; this result would also be indicative of the emergence of a shared mutual understanding of an application type, which could be called, for example, "legal data and content curation system".
- Finally, the results obtained in the one, two or three workshops will be translated into requirements for the design and implementation of the Curation Workflow Manager.

Such a user-centred design approach allows the inclusion of the users (in our case, three use case partners) in the requirements gathering process, because they contribute to the initial definition of requirements through the survey, and also in its concretization through the workshops.

3.1. Survey

The first step of the requirements gathering process is a survey that will help us to define the general needs of the use case partners (and their clients). The survey is divided into several parts described in the next sections.

3.1.1. Non-functional Requirements

The non-functional requirements part of the survey has the goal to sketch the most general and abstract needs the pilot use case partners have in relation to the project and the overall platform (see Question 1. in Section 1.).

1. Please describe, as specifically as possible, your use case (or use cases): what kind of functionality or processing capabilities do you want to realise or achieve with the help of the Lynx platform?
2. What kind of devices do you work with predominantly? (Desktops/laptops, touch-interface devices, speech interfaces etc.)
3. Do you plan to integrate the Lynx platform into existing in-house systems and graphical user interfaces (GUIs)?

If the answer to question 3 is "Yes", please also reply to questions 4 and 5:

4. Please specify the system into which you want to integrate Lynx. Please provide screenshots or screencasts of the system.

5. Do you currently use a stand-alone application with a GUI or web-based GUI?

If the answer to question 3 is "No", please also reply to questions 6 and 7:

6. How are you planning to use the services developed in Lynx? (REST API calls, Web services, Web browser, Mobile phone/tablet applications, Other)

7. Would your preference be to develop a new (web-based) GUI to connect to the Lynx services or would you prefer some other way?

3.1.2. Actual Usage of Automatic Processing

This part of the survey is intended to analyse the current usage of automatic processing techniques and tools inside the use case partners environments and their customers (legal firms).

8. How do you analyse or process legal documents in your company? (For example, with the help of human experts, fully automatically, semi-automatically etc.? Please be as specific and descriptive as possible.)

9. Do you use automatic solutions and tools for analysing and processing legal documents in your company? If yes, which ones?

10. What kind of documents from the legal domain (or your use case domain) do you work with (official law texts, letters, case law, EU regulations and directives, client specifications etc.)?

11. If you already use software for processing legal documents, please provide screenshots or screencasts of your software/GUIs.

12. In terms of use cases and workflows, please specify all (or a representative set of) typical workflows that you use in-house (e. g., types of documents, types of analysis, types of processing, types or approaches of producing new content, etc.).

The questions in this part of the survey are rather abstract and general, because we need to get an overview of the workflows that the use case partners currently use, without paying too much attention to the implementation or concrete details (with which we will deal in later steps in the project).

3.1.3. Users and Profiles

Although this part of the survey does not have many questions, they are important for the development of the Curation Workflow Manager, because depending on the amount and type of users that can use the platform (workflows) the whole management implementation has to be adapted.

13. What types of users are going to use Lynx services (e. g., JavaScript developers, lawyers, knowledge workers, customers, etc.)?

14. Do you need a multi-user solution?

15. Do you need authentication (login/password)?

16. Do you need access control lists with different roles and different permissions?

3.1.4. Data Sets

This part of the survey is more concrete and tries to get a better understanding about the concrete datasets needs in every use case (addressing Questions 2. and 4. in Section 1.). The idea is to determine which datasets are needed and in which format for every use case.

17. What kind of reference materials or reference data sets do you use on a regular basis?

18. Which online data sets or reference materials would help you in your daily work?

19. File Formats: Which are the formats of files that you want to process with Lynx? Do you want the same file format in the request you send to Lynx as well as in the responses you get back from Lynx?

3.1.5. Common Services

This part of the survey is more concrete and tries to get a better understanding about the concrete services needs in every use case. The idea is to determine those services that are needed overall and those that are specific to only one concrete case. This section addresses Questions 2., 3. and 4. from Section 1..

20. Do you need a tool that can identify and highlight named entities (persons, locations, organizations, etc.) in legal documents? For example, this could result in a colour-based highlighting of person, location, organisation names in documents or the filtering of document collections based on the names contained in them.

21. Do you need a tool that can identify and highlight time expressions and normalize them? Such a function could enable a timeline view of a large document collection, for example, of a series of letters or correspondence.

22. Do you need a tool that can identify and highlight geographical information related to locations in legal documents? For example, the output of such a function could be an interactive map containing all documents or content of the documents.

23. Do you need a tool that can identify and highlight events (or other types of important keywords) in legal documents? For example, the output of such a function could be a list of events (words, phrases, expressions, etc.) that require some kind of action or reaction from the reader.

24. Do you need a tool that can identify relations between entities (some judge is related to a criminal because they are involved in a court case) in legal documents? For example, the output of such a function could result in capabilities for searching documents containing relations through certain entities.

25. Do you need a tool that can identify specific domain terminology (legal terms, oil & gas related terms, etc.) in legal documents?

26. Do you need a tool that can recognize citations, references and relations between legal documents? For example, the output of such a function could be an interactive graph display showing the relations between all the documents of a court case or piece of legislation.

27. Do you need a tool that can disambiguate the sense of a term determining if it is referring to labour law (as an example) or any other domain in legal documents? For example, the output of such a function could be used for better determining concrete topics the document is talking about.

28. Do you need a tool that can translate legal documents to other languages (if yes, which languages and language pairs?)?

29. Do you need a tool that can summarise documents or sets of documents in the legal domain?

30. Do you need a tool that can search through collections of legal documents?

31. Do you need a tool that can recommend other legal documents related to a certain task?

32. Do you need a tool that can alert you about changes in existing legal documents or the appearance of new legal documents?

33. Do you need a tool that can determine the main topic of a legal document or part of a document (paragraphs, etc.)? For example, the output of such a function could help in searching documents for certain legislations.

34. Do you need a tool that can determine the main type of a legal document (e. g., letter, law, contract, technical report, case report etc.)? For example, the output of such a function could help further process and visualise a large and heterogeneous set of documents.

35. Do you want to combine several automatic processing steps? For example: When you get a document, the first thing you do is to translate (if it is in a language other than English), then you read it to learn which people are mentioned (locations and time expressions are also important but first are people). After that you focus on the references of other laws and finally you try to identify arguments and events.

3.1.6. Additional Requirements

The last part of the survey is an open question for including any information that is missing in the previous questions and that the use case partners what to include.

36. Please write down any additional requirements you may have that are not covered by the questions above.

3.2. Workshops

Once the survey has been circulated to the pilot use case partners and they have filled it in, we will analyze them to define the first sets of requirements. With these we will be in a good position to define and plan the workshops in which we will concretize (clean and filter) the requirements with the partners.

The development of the workshops depend directly on the results obtained from the surveys. If the implementation of a completely new and redesigned GUI is considered important (through Question 5), the workshop will be established as a graphical design workshop where the main focus will be put on the generation of mockups and wireframes of the new interface. Here, the output of the workshop will not only be a list of requirements, but also a set of mockups of the new interface. Depending on the results of the survey, it has to be decided if there will be only one workshop with the three use case partners, developing a common GUI as well as individual solutions, or if it is better to have independent workshops, one with each use case partner and then extrapolating a common interface. If the surveys reveal, on the other hand, that an integration of the new components, services and workflows in existing systems is needed, three individual workshops will be organised. These will be focused on the study and analysis of the currently used technologies and how the users interact with them, as well as how the future workflows can be integrated into the current processes.

4. Summary and Conclusion

We define the concept of the Curation Workflow Manager (CWM), which refers to the management of specifying different curation workflows in an adaptive platform. The definition of workflows is a complex task that requires a close collaboration among all involved stakeholders by means of requirements gathering processes. We apply a hybrid approach that consists of a requirements gathering survey, together with face-to-face workshops with the pilot use case partners. The survey has been design with the main goal of gathering concrete information about the three pilot use cases in LYNX. The first part collects general information about the intended use of LYNX technologies. The second part is designed to learn more details about the use case partners' current workflows, in addition to determining if automatic processes are used. The survey also includes questions on the intended users of the system and regarding the functional requirements. The last part collects information on the necessary infrastructure for each of the use cases. We are still in the process of design the Curation Workflow Manager. This paper includes the final version of the survey. First results of the requirements gathering phase will be presented at the workshop and in follow-up publications.

5. Acknowledgements

The project LYNX has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 780602. More information is available online at http://www.lynx-project.eu.

6. Bibliographical References

Baumgartner, W. A., Cohen, K. B., and Hunter, L. (2008). An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *Journal of biomedical discovery and collaboration*, 3(1):1.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, pages 19–10.

Bourgonje, P., Schneider, J. M., Rehm, G., and Sasaki, F. (2016). Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In Aldo Gangemi et al., editors, *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 13–16, Edinburgh, UK, September. The Association for Computational Linguistics.

Bourgonje, P., Schneider, J. M., and Rehm, G. (2017). Domain-specific Entity Spotting: Curation Technologies for Digital Humanities and Text Analytics. In Nils Reiter et al., editors, *CUTE Workshop 2017 – CRETA Unshared Task zu Entitätenreferenzen. Workshop bei DHd2017*, Berne, Switzerland, February.

BPM, B. P. M. (2009). Glossary. Technical report, Center of Excellence (CoE).

Chard, K., Russell, M., Lussier, Y. A., Mendonça, E. A., and Silverstein, J. C. (2011). A cloud-based approach to medical NLP. In *AMIA Annual Symposium proceedings*, volume 2011, page 207. American Medical Informatics Association.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust NLP tools and applications. In *ACL*, pages 168–175.

de Castilho, R. E. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11.

Fricker, S. A., Grau, R., and Zwingli, A., (2015). *Requirements Engineering: Best Practice*, pages 25–46. Springer International Publishing, Cham.

Hahn, U., Buyko, E., Landefeld, R., et al. (2008). Language Resources and Evaluation Workshop, Towards Enhanced Interoperability Large HLT System: UIMA NLP. *Marrakech, Morocco*, pages 1–8.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34(suppl_2):W729–W732.

Islam, M., Huang, A. K., Battisha, M., Chiang, M., Srinivasan, S., Peters, C., Neumann, A., and Abdelnur, A. (2012). Oozie: Towards a Scalable Workflow Management System for Hadoop. In *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*, SWEET '12, pages 4:1–4:10, New York, NY, USA. ACM.

Kano, Y., Baumgartner Jr, W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L., and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, 25(15):1997–1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L. E., Ananiadou, S., and Tsujii, J. (2011). U-Compare: A modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3):11–1.

Köster, J. and Rahmann, S. (2012). Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

Perovšek, M., Kranjc, J., Erjavec, T., Cestnik, B., and Lavrač, N. (2016). TextFlows: A visual programming platform for text mining and natural language processing. *Science of Computer Programming*, 121:128–152.

Rak, R., Rowley, A., Black, W., and Ananiadou, S. (2012). Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database*, 2012:bas010.

Rehm, G., He, J., Schneider, J. M., Nehring, J., and Quantz, J. (2017a). Designing User Interfaces for Curation Technologies. In Sakae Yamamoto, editor, *Human Interface and the Management of Information: Information, Knowledge and Interaction Design, 19th International Conference, HCI International 2017*, number 10273 in Lecture Notes in Computer Science (LNCS), pages 388–406, Vancouver, Canada, July. Springer. Part I.

Rehm, G., Schneider, J. M., Bourgonje, P., Srivastava, A., Nehring, J., Berger, A., König, L., Räuchle, S., and Gerth, J. (2017b). Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters. In Tommaso Caselli, et al., editors, *Proceedings of the Events and Stories in the News Workshop*, pages 42–51, Vancouver, Canada, August. Association for Computational Linguistics. Co-located with ACL 2017.

Rehm, G., Schneider, J. M., Bourgonje, P., Srivastava, A., Fricke, R., Thomsen, J., He, J., Quantz, J., Berger, A., König, L., Räuchle, S., Gerth, J., and Wabnitz, D. (2018). Different Types of Automated and Semi-Automated Semantic Storytelling: Curation Technologies for Different Sectors. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 232–247, Cham, Switzerland, January. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Schneider, J. M., Bourgonje, P., and Rehm, G. (2017). Towards User Interfaces for Semantic Storytelling. In Sakae Yamamoto, editor, *Human Interface and the Management of Information: Information, Knowledge and Interaction Design, 19th International Conference, HCI International 2017*, number 10274 in Lecture Notes in Computer Science (LNCS), pages 403–421, Vancouver, Canada, July. Springer. Part II.

Unertl, Kim M and Novak, Laurie L and Johnson, Kevin B and Lorenzi, Nancy M. (2010). Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review. *Journal of the American Medical Informatics Association*, 17(3):265–273.

Van Der Aalst, W. and Van Hee, K. M. (2004). *Workflow management: models, methods, and systems*. MIT press.

Event Extraction from Legal Documents in Spanish

Gerardo Sierra[1], Gemma Bel-Enguix[1], Guillermo López-Velarde[1], Ricardo Saucedo[2], Lucía Rivera[1]

[1] Universidad Nacional Autónoma de México, [2]Avoquate

[1]Instituto de Ingeniería, UNAM, Circuito Escolar S/N Instituto de, Ingeniería, Cd. Universitaria, 04510, Ciudad de México, México, [2]Avenida Oaxaca 31, colonia Roma Norte, delegación Cuauhtémoc, 06700, Ciudad de México, México

[1]{GSierraM, Gbele, GLopezVelardeG, LRiveraV}@iingen.unam.mx, [2]ricardos@avoquatemaker.com

Abstract

This work is part of a more general project aiming to design a tool that can help lawyers to find the information they need for litigation in a fast and efficient way. The resource is being designed for Spanish, a language that has a scarceness of Natural Language applications for legal coding, and is tested in 300 documents, mainly writs of 'amparo', a legal procedure to protect human rights, by means a judicial review of governmental action. These documents have been freely downloaded from the Mexican Instituto Federal de Telecomunicaciones. The system, implemented in Python, will include modules to perform several tasks, like automatic classification, Named Entities identification, law detection, structure summarization, and event extraction. This article is focused in one of the most complex parts of the development, event extraction. The algorithm works linking dates with events in the texts. These events are reduced to a list of verbs that have been reported as the most meaningful in this type of texts. For every verb-event, a list of pieces of information will be retrieved: 'who', 'what', 'to whom' and 'where'.

Keywords: legal documents, event extraction, natural language patterns

1. Introduction and Motivation

This paper presents a system for extracting events from legal texts in Mexican Spanish. This is part of a more general project aiming building a tool that can help lawyers to find the information they need in a fast and efficient way.

Our research has been focused in finding patterns for Spanish sentence structures that are used in legal documents. We have worked with 300 documents downloaded from the Mexican 'Instituto Federal de Telecomunicaciones' (IFT). This organization has an open webpage[1] where its resolutions can be accessed.

This article explains the methodology of the system, and its initial performance when trying to automatically detect events and its related date.

2. Previous Work

Lawyers need the processing and study of large quantities of documents as a part of their everyday life. Not having tools available for automatically obtaining the required data from texts, they perform these tasks manually, in what is an expensive and time-consuming activity. Computational linguistics can help lawyers to automatically process the documents they need. Many aspects can be taken into account when dealing with litigation documentation, from consulting laws to getting information of related trials. In what refers to laws, vlex[2] provides an extensive coverage of legislation, including Mexico. This resource offers also links to other laws the text refers to. As for documents generated by litigations, there are several attempts to build databases that can relate some documents with others.

However, it is necessary to have tools able to search in these documents for the information that a lawyer can need in the professional activity. The field of ICT applied to Law has created the area of legal technologies. Sartor et al. (2008) summarize the major types of resources related to legal technologies: legal information search, electronic data discovery, web-based communications, collaborative tools, Metadata and XML Technologies and Technologies in Courtrooms and Judicial Offices. In the last years, however, the area of legal text processing and information extraction, more closely related to Natural Language Processing, has been developed (Francesconi et al., 2010).

A key topic in automatic processing of legal texts is the identification of people and organizations related to a legal case. This is very much related to entity recognition, but must be focused in the fact that these entities need to have a given role in the legal case. In this area, there are several contributions. Dozier et al. (2010) create a hybrid system for named entities recognition and resolution in legal texts, while Quaresma and Gonçalves (2010) use machine learning techniques for solving the same problem. Kumaran & Alan (2004) design a system for NE recognition for new event detection, but it is not related to legal texts. A collection of resources that can be used to deal with legal texts can be found in the document Collection of state-of-the-art NLP tools for processing of legal text, from the project MIREL[3].

The Automatic Context Extraction (ACE)[4] evaluation defines an event as 'something that happens or leads to some change of state' (Nguyen et al., 2016). Meanwhile, Pustejovsky et al. (2002) define it as those expressions into a narrative that can be ordered temporary. This idea was the basis for the organization of TempEval shared tasks (UzZaman et al., 2013), that have helped to the development and testing of of different systems for event extraction and ordering. The area has been a trending topic in text mining. Hogenboom et al. (2011) distinguish three main approaches to the problem: a) data-driven, that try to convert data to knowledge by means to statistics, machine learning, etc.; b) knowledge-driven approaches, that are mainly pattern-based; and c) hybrid, that combine the other models.

Knowledge-driven methods are based on linguistic and lexicographic knowledge. Information is mined using

[1]http://apps.ift.org.mx/cumplimientoStp/secured/adminficum.faces

[2] https://app.vlex.com/

[3] MIREL: Mining and Reasoning with Legal Texts: http://www.mirelproject.eu

[4] https://www.ldc.upenn.edu/collaborations/past-projects/ace

semantic or syntactic patterns. Some examples are Nishihara et al. (2009) and Aone & Ramos-Santacruz (2000). The system Evita to extract events focus on verbs, nouns, nominal phrases and adjectival phrases (Saurí et al., 2005). Other works on event processing (Mani et al., 2003; Filatova y Hovy, 2001) use tools like CLAUSE-IT or CONTEX (Hermjakob y Mooney, 1997) to identify syntactic structures.

Some authors have developed methodologies to extract events from specialized domains. Yakushiji et al (2001) apply this method in the biomedical domain, while Li et al. (2002) work in the financial area and Cohen et al. (2009) focus in biology. As for legal texts event extraction, there is an interesting contribution with English documents (Lagos et al., 2010), based in a semi-automatic approach that integrates two main components: information extraction, and knowledge integration.

Our work fits in the area of knowledge-driven methods, and uses well-known common patterns from legal texts. However, the area is not enough developed in Spanish, and this work presents a small advance in the concrete space of Mexican legal system.

3. Legal Language and Patterns

In order to achieve consistency, validity, completeness and soundness, legal texts are subject to certain constraints, both with respect to content and form. They follow a rigid structural format. Legal writing uses a lot of legal terminology and scholarly words, but specially some linguistic patterns. Danet (1985) describes some legal English features, such as archaic expressions, doublets, unusual prepositional phrases, passive constructions, long sentences and syntactic complexity. Collectively, these features are often called legalese.

For example, among archaic expressions found in legal Spanish documents, there is a frequent use of the expression hereinafter (en lo sucesivo).

...se creó el Instituto Federal de Telecomunicaciones (en lo sucesivo, el "Instituto").

Knowing this type of expressions can be very important for automatic information extraction in legal texts. Being aware that 'Instituto' is a short name, or alias, to name the 'Instituto Federal de Telecomunicaciones' allows the identification of the actant intervening in the event.

Likewise, knowing the syntactic complexity of legal language is very useful to differentiate the relevant information in the description of the event. The use of large sentences and the insertion of appositions is frequent in this type of texts.

Example 1

El 13 de diciembre de 2006, de conformidad con los artículos 13 de la LFT, 16 y 21 de la LFRTV, la COFETEL otorgó a favor del Concesionario, el refrendo de la Concesión para operar y explotar el canal 7. (P_IFT_111215_577_Acc.docx)[5]

In the description of the event of Example 1, several pieces of information can be extracted. First, the date (13/12/2006). The second element is what was done (se otorgó el refrendo [the endorsement was granted]). Another item is who did this (COFETEL) and to whom (el Concesionario [the dealer]). To get all this information several steps have to be performed: a) discarding non-relevant information. In Example 1, it is the apposition (de conformidad con los artículos 13 de la LFT, 16 y 21 de la LFRTV); b) Identifying the 'Who', 'What' and 'to Whom' of the sentence, which most of times are the subject, object and indirect object, respectively.

4. Methodology

Although our goal is the application of the methodology to any type of legal text, so far we have been working with a collection of 300 texts downloaded from the IFT of Mexico, which are freely available on their website. Most of them are what is called writ of 'amparo' in mexican legislation. 'Amparo' is a legal procedure to protect human rights, by means a judicial review of governmental action.

The description of events usually follows a regular pattern involving at least two elements: the action, determined by the main verb, and the date on which the event occurred. In this sense, an analysis was made of the verbs that occur in the writings of amparo, as well as the direct objects of each verb. The description is given below.

4.1 Pre-processing

The first steps in the processing of the corpus are:

a) Change the files format from .docx to .txt.

b) Replace 17 text patterns to help FreeLing 4.0[6] make a better PoS tagging. These patterns can be divided into 3 categories: misspells, conjunctions and business entities types.

Examples of each category can be found in the Table 1:

Replace	With
Con cesiones	Concesiones
y Transportes	Y_Transportes
, S.A. de C.V.	Sadecv

Table 1: Pattern replacement in pre-processing

For example, the word 'Con cesiones' is recurrently misspelled, as it should be 'Concesiones'.

The entity 'Secretaría de Comunicaciones y Transportes' is wrongly tagged as follows: Secretaría de Comunicaciones (NP00000), y (CC), Transportes (NP00000). So, we replace 'y Transportes' with 'Y_Transportes' in order to get the whole entity tagged as NP00000.

Finally, in México there are different types of business entities that can be legally constituted, which names are always referred when a company name is mentioned, for example, the entity 'Telefonía Inalámbrica del Norte, S.A de C.V.'. In this case, the entity is not tagged as NP00000 because the type is referred after a comma. So, we replace ', S.A de C.V.' with 'Sadecv' to avoid further confusions. We also replace the instances where there is no comma

[5] This reference is the name of the document that can be downloaded from the webpage of the IFT.

[6] http://nlp.lsi.upc.edu/freeling/node/1

that separates the business entity type with the name of the company, to homogenize all business entities.
c) PoS tagging by means of FreeLing 4.0. Made with the default Spanish configuration file. In this step, FreeLing also identifies dates, assigning the tag 'W'.
d) Identification of Named Entities (NE). With what the system obtains here, a table is made that will later be modified, if necessary, during the next steps of the whole system. In the meantime, this table serves as basis to detect actants in the events.
All preprocessing is implemented in python. Step d) does not rely only on freeling to identify named entities. The table is built using another rule-based system we have developed for writs of amparo.

4.2 Verbs and dates

Within the investigations that address event detection in text, we found that most of them try to find words, phrases or indicators that establish the point in time where the events happen. That is, they seek to find what linguistic elements are used to express moments or successions, so the computers can use them in a standardized manner. One of the clearest ways in which a point in time or an interval can be represented is by identifying dates.
Every document in our corpus has the date of release, the date of submission and, sometimes, the signature date. Finding these elements is not the goal of the paper, but detecting the ones that are linked to an event in the text of the resolution.
In order to do this, the procedure starts from the idea that, in this type of document, every event is related to a date, and every event is characterized by a main verb that represents the action that is being made. So, all the dates that do not contain such verb, are not taken into account.
Additionally, every event has some actants related to the verb, which correspond to the Named Entities that have to be extracted in the pre-processing task d).
The dates are tagged by FreeLing 4.0 in the pre-processing task c).
Regarding the verbs, we found, by manually analyzing the data, that in the type of documents that we are dealing with, writs of 'amparo' of the IFT, almost every event is correlated to one of the following verbs: 'emitir' [release, issue], 'otorgar' [grant], 'presentar' [submit], 'publicar' [publish], 'solicitar' [request].
The information required for every event in the document is the one in Table 2:

	Who	What	To Whom	Where
emitir	YES	YES	NO	NO
otorgar	YES	YES	YES	NO
presentar	YES	YES	YES	NO
publicar	YES	YES	NO	YES
solicitar	YES	YES	YES	NO

Table 2: Main information items for each one of the verbs that configure the events

In the sequel, the main patterns that have been used for every one of the elements of information are discussed.

4.3 Quién [Who]

If the verb is in active voice, 'Who' is the subject, and it is at the left. This has to be a NP, present in the table of NE of the system.
If the verb is in passive voice, 'Who' is located at the right side, it must be a NP in the table of NE, and it fits into the pattern: 'por + NP'.
In legal texts, some other more elaborated models for 'Who' can be found, both at left or at right of the verb. In 'Who' patterns, the NP is always a NE.
Frequent structures are the ones in which some NPs are explained by other NPs, being both the 'Who' of the event, the second NP can be delimited by colons, or not, and it is an NE. Some common patterns for this structure are:

(1) <[NP] + (,) + 'representante legal de' + [NP](,)>
<[NP]+ (,) + legal representative of + [NP](,)>[7]
(2) <[NP] + (,) + 'mediante' + [NP](,)>
<[NP] + (,) + through+ [NP](,)>
(3) <[NP] + (,) + 'por medio de' + [NP](,)>
<[NP] + (,) + through + [NP](,)>
(4) <[NP] + (,) + 'a través de' + [NP](,)>
<[NP] + (,) + through + [NP](,)>
(5) <[NP] + (,) + [alias]>

Example 2 illustrates pattern (1), where the 'Who' is 'el representante legal de Axtel'.

Example 2
El 16 de octubre de 2009 el representante legal de Axtel presentó ante la Comisión Federal de Telecomunicaciones el escrito No. 321-2009 mediante el cual solicita la intervención de este órgano a efecto de que resuelva los términos condiciones y tarifas aplicables a partir del 10 de enero de 2010 que no ha podido convenir con Telmex y Telnor para la interconexión de sus respectivas redes públicas de telecomunicaciones.
(P_IFT_140410_191.docx)

Example 3 shows more specifically pattern (4). The 'Who' is 'Unidad de Competencia Económica', but this entity is not the one that issued the trade, but another one on its behalf, the 'Dirección General de Concentraciones y Concesiones'.

Example 3
Con fecha 14 de mayo de 2015 la Unidad de Competencia Económica a través de la Dirección General de Concentraciones y Concesiones emitió el oficio IFT mediante el cual remite la opinión correspondiente a la Solicitud de Prórroga. (P_IFT_170316_125_Acc.docx)

Finally, the 'Who' piece can follow the pattern (5), as in 'Telefonos de México, Telmex', where Telmex is an alias

[7] Translations to English are orientative. 'mediante', 'por medio de' and 'a través de' can be roughly translated to 'through'. And they mean that a person is doing something instead of another person who she/he represents.

that works usually instead of 'Teléfonos de México, S.A.B. de C.V.'.
These patterns are not all that can define 'Who' in a legal text, but the ones that can capture almost every structure in the sub-genre of writs of amparo.

4.4 Qué [What]

If the verb is in active voice, 'What' is usually immediately after the verb, at its right. If the verb is in passive voice, the 'What' is at left.
It is a NP, VMN, VMP or VMS. In the Example 2, the word 'escrito' has the form of a VMP.
In the verb 'publicar', the 'What' is usually found in quotes, as shown in Example 4, where 'Decreto por el que se expiden la Ley Federal de Telecomunicaciones y Radiodifusión, (...)' is marked as 'What'.

4.5 Dónde [Where]

After the analysis of the documents, we found that only the verb 'publicar' is expected to have this information. To find it, we use the pattern <'en' + NP>, which we observed to be the most common for this verb. In Example 4, 'Diario Oficial de la Federación' fits said pattern.

Example 4
El 14 de julio de 2014 se publicó en el Diario Oficial de la Federación el "Decreto por el que se expiden la Ley Federal de Telecomunicaciones y Radiodifusión y la Ley del Sistema Público de Radiodifusión del Estado Mexicano; y se reforman adicionan y derogan diversas disposiciones en materia de telecomunicaciones y radiodifusión" mismo que entró en vigor el 13 de agosto de 2014. (P_IFT_170216_57_Acc.docx)

4.6 A quién [To whom]

To find the phrase that stands for 'to Whom', some common patterns are:

(1) <'a' + NP>
<*to + NP*>
(2) <'ante' + NP>
<*before + NP*>
(3) <'en favor de' + NP>
<*in favor of + NP*>
(4) <'a quien' + NP>
<*to whom + NP*>

This NP must be located immediately after one of the verbs that are considered, or at least they do not have to have any other verb between both elements.
In Example 5 'C. Ricardo León Garza Limón' is marked as 'to Whom', because it fits pattern (3) and the NP comes after the main verb 'otorgó', without any other verbs in between.

Example 5
El 18 de octubre de 2005 la Secretaría de Comunicaciones Y Transportes (la "Secretaría") otorgó en favor de el C. Ricardo León Garza Limón un título de concesión.

5. Discussion and Future Work

This is a work in progress that aims at finding patterns in Legal Language in Mexican Spanish in order to extract events in writs of 'amparo'.
The application has retrieved good results so far, but the system must be improved in several ways, for: a) designing a system capable to obtain every pattern for each element of information, due to the ones that have been implemented so far do not cover every possible case, but only the more general ones; b) taking into account juridic expressions in non-Mexican Spanish; c) being extended to other specific areas of litigation, which may result in a wider variety of verbs that can define new types of events.
An important area that should be improved is evaluation. So far, the only way to do it is manually by humans.
Also, in the future we aim to implement a machine learning algorithm trained with manually annotated data, so we can compare this rule-based system to the supervised learning algorithm and figure out which approach is better in the long run, thinking that this information may someday be part of a bigger system.
From this seminal design we plan to build a system that can efficiently extract every piece of information lawyers can need from a legal text, and design friendly systems that can truly help in the court.

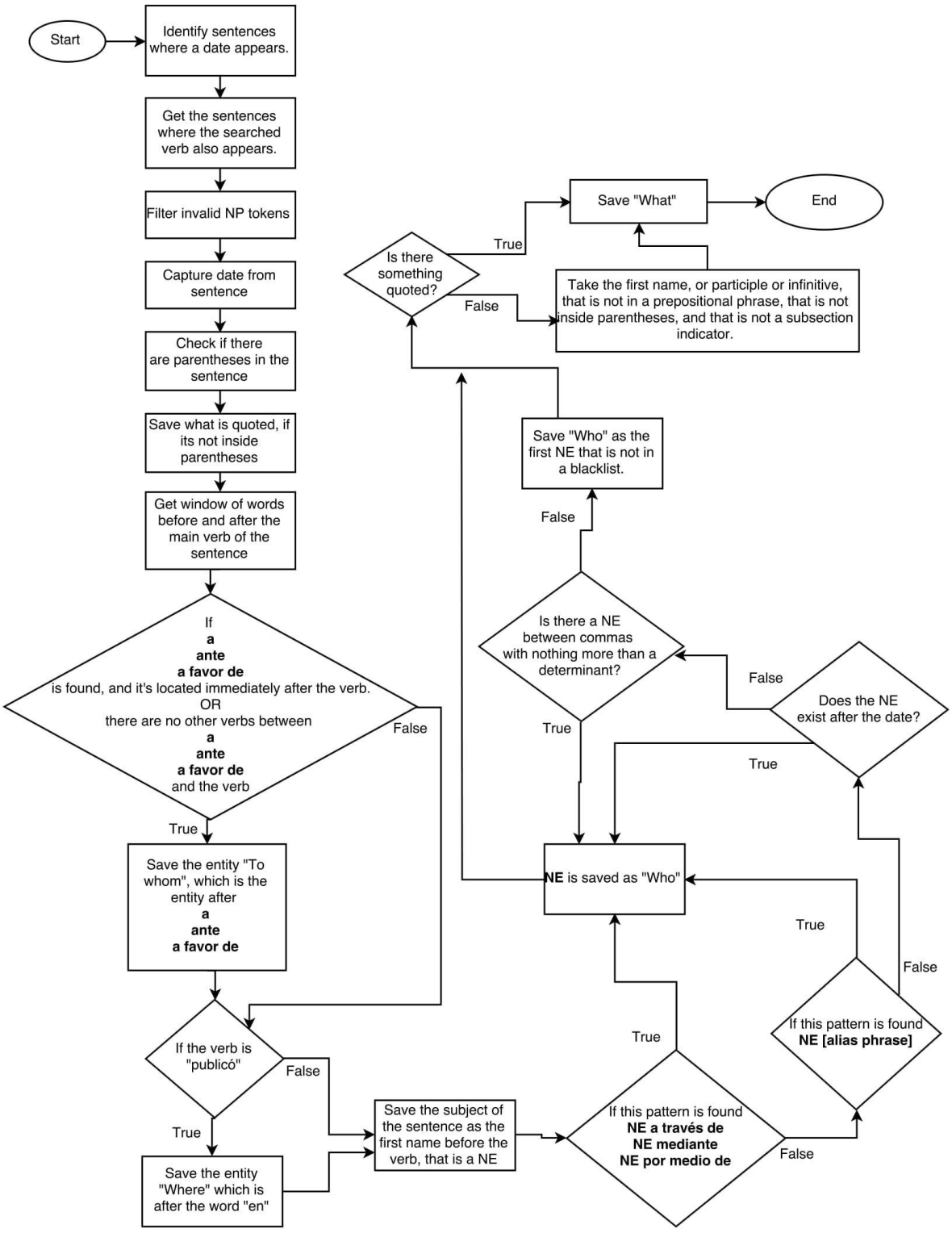


Figure 1: Scheme showing the steps used to implement the event identification system

6. Bibliographical References

Aone, C., Ramos-Santacruz, M. (2000). REES: A Large-Scale Relation and Event Extraction System. In: 6th Applied Natural Language Processing Conference (ANLP 2000): 76–83. Association for Computational Linguistics.

Cohen, K.B., Verspoor, K., Johnson, H.L., Roeder, C., Ogren, P.V., Baumgartner, Jr., W.A., White, E., Tipney, H., Hunter, L. (2009). High-Precision Biological Event Extraction with a Concept Recognizer. In: Workshop on BioNLP: Shared Task collocated with the NAACL-HLT 2009 Meeting. pp. 50–58. Association for Computational Linguistics (2009).

Danet, B. (1985). "Legal Discourse". In Teun A. Van Dijk (ed.) Handbook of Discourse Analysis. Vol. 1, 237 – 291 . London : Academic Press.

Dozier C., Kondadadi R., Light M., Vachher A., Veeramachaneni S. and Wudali R. (2010). Named Entity Recognition and Resolution in Legal Text. In Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds) *Semantic Processing of Legal Texts. Lecture Notes in Computer Science*, vol 6036. Springer, Berlin, Heidelberg.

Filatova, E. and Hovy, E. (2001). Assigning time-stamps to event-clauses. In Proceedings of the workshop on Temporal and spatial information processing-Volume 13, p. 13.

Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds) (2010). *Semantic Processing of Legal Texts. Lecture Notes in Computer Science*, vol 6036. Springer, Berlin, Heidelberg.

Hermjakob, U. y Mooney, R.J. (1997). Learning parse and translation decisions from examples with rich context. En Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, págs. 482–489.

Hogenboom, F., Frasincar, F., Kaymak, U, de Jong, F. (2011), An Overview of Event Extraction from Text. In Workshop on Detection, Representation and Exploitation of Events in the Semantic Web (DeRiVE 2011), vol 799, pags 48-57, CEUR Workshop Proceedings.

Kumaran, G. and Allan, J. (2004). Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '04). ACM, New York, NY, USA, 297-304. DOI=http://dx.doi.org/10.1145/1008992.100904.

Lagos, N., Segond, F., Castellani, S. and O'Neill, J. (2010). Event Extraction for Legal Case Building and Reasoning. In Zhongzhi Shi, Sunil Vadera, Agnar Aamodt, David Leake. Intelligent Information Processing V, 340, Springer, pages 92-101, 2010, IFIP Advances in Information and Communication Technology, 978-3-642-16326-5. <10.1007/978-3-642-16327-2 14>. <hal-01055067>

Li, F., Sheng, H., Zhang, D. (2002) Event Pattern Discovery from the Stock Market Bulletin. In: 5th International Conference on Discovery Science (DS 2002). Lecture Notes in Computer Science, vol. 2534, pp. 35–49. Springer-Verlag Berlin Heidelberg.

Mani, I., Schiffman, B., y Zhang, J. (2003). Inferring temporal ordering of events in news. En Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003 – short papers - Volume 2, págs. 55–57.

Morris, F. J. (2005). E-Discovery: Best practices for employment lawyers. what support do you need? how do you work with E-Discovery experts. Current Developments in Employment Law, ALI-ABA, Santa Fe, NM.

Nguyen, T.H., Cho, K. and Grishman, R. (2016). Joint Event Extraction via Recurrent Neural Networks, Proceedings of NAACL-HLT 2016, pages 300–309.

Nishihara, Y., Sato, K., Sunayama, W. (2009). Event Extraction and Visualization for Obtaining Personal Experiences from Blogs. In: Symposium on Human Interface 2009 on Human Interface and the Management of Information. Information and Interaction. Part II. Lecture Notes in Computer Science, vol. 5618: 315–324. Springer-Verlag Berlin Heidelberg.

Pustejovsky, J., Saurí, R., Setzer, A., Gaizauskas, R., and Ingria, B. (2002). TimeML annotation guidelines. TERQAS Annotation Working Group 23.

Quaresma, P., & Gonçalves, T. (2010). Using linguistic information and machine learning techniques to identify entities from juridical documents. In Semantic Processing of Legal Texts (pp. 44-59). Springer Berlin Heidelberg.

Sartor, G., Casanovas, P., Casellas, N., Rubino, R. (2008). Computable models of the law and ICT: State of the art and trends in european research. In: Casanovas, P., Sartor, G., Casellas, N., Rubino, R. (eds.) Computable Models of the Law. LNCS (LNAI), vol. 4884, pp. 1–20. Springer, Heidelberg.

Saurí, R., Knippen, R., Verhagen, M., y Pustejovsky, J. (2005). Evita: a robust event recognizer for QA systems. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, págs. 700–707.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Vol. 2, pp. 1-9).

Yakushiji, A., Tateisi, Y., Miyao, Y. (2001). Event Extraction from Biomedical Papers using a Full Parser. In: 6th Pacific Symposium on Biocomputing: 408–419 (2001).

Legal text processing within the MIREL project

Milagro Teruel*, **Cristian Cardellino***, **Fernando Cardellino***, **Laura Alonso Alemany***, **Serena Villata**†
*Universidad Nacional de Córdoba Argentina
INRIA & Université Côte d'Azur CNRS France

Abstract

We present the roadmap and advances in the area of Information Extraction from legal texts within the EU-funded MIREL project (MIning and REasoning with Legal texts). We describe the resources and tools we have developed for Natural Language Processing in the legal domain, i.e., annotated corpora and automated classifiers for Named Entity Recognition and Linking and Argument Mining. Our final objective is to identify arguments, their content and the relations between them in legal text, with a proof-of-concept in judgments of the European Court of Human Rights (ECHR), to finally support reasoning tasks over mined argumentative structures. This representation will arguably be useful for applications like a reading aid, enhanced information retrieval, structured summarization, intelligent search engines or information extraction. All tools and resources are available at `https://github.com/PLN-FaMAF/legal-ontology-population` and `https://github.com/PLN-FaMAF/ArgumentMiningECHR`.

Keywords: Argument Mining, Named Entity Recognition, Classification and Linking, Legal Information Extraction

1. Introduction and Motivation

Automated legal text processing is becoming more and more relevant within legal practice. According to the MIT Technology Review, the U.S. Consultancy group McKinsey estimates that 22% of a lawyer's job and 35% of a law clerk's job can be automated (Winick, 2017), for example:

> "JPMorgan announced earlier this year that it is using software called Contract Intelligence, or COIN, which can in seconds perform document review tasks that took legal aides 360,000 hours." "CaseMine, a legal technology company based in India, builds on document discovery software with what it calls its "virtual associate," CaseIQ. The system takes an uploaded brief and suggests changes to make it more authoritative, while providing additional documents that can strengthen a lawyer's arguments."
>
> (Winick, 2017)

Natural Language Processing (NLP) tools have the capabilities scan huge amounts of legal documents, identify portions relevant to a given case and even present them in an orderly manner for a lawyer needs to craft a case, more quickly and more exhaustively than humans given the huge amount of data to process. In case law, if law practitioners are provided with relevant cases when they are building their arguments for a new case, they could be more liable to produce a sounder argumentation. It is also to be expected that cases are resolved more definitely if compelling jurisprudence is provided, even at an early stage in the judicial process. More and more technological solutions are being developed in this line, which shows the feasibility and utility of this line of work.

One of the objectives of the MIREL project[1] is to develop tools for MIning and REasoning with Legal texts, with the aim of translating these legal texts into formal representations that can be used for querying norms, compliance checking, and decision support. Open-source tools and resources are also very important to provide equality in the access to the law. However, developing such tools is costly. Tools are usually trained with examples that have been manually analyzed and annotated by a domain expert, so we aim to reduce the cost of developing such tools by taking advantage of existing annotated resources.

In this paper, we present our roadmap and advances to develop such tools, working in two main areas: Named Entity Recognition, Classification and Linking (NERC and NEL) and Argument Mining (Lippi and Torroni, 2016). For each of these two areas, we both built annotated datasets following precise guidelines, and experimented supervised and unsupervised learning methods. More precisely, we have built a tool for NERC and NEL in the legal domain by exploiting the Wikipedia as an annotated corpus. To retrieve the relevant portion of the Wikipedia, we have established a mapping between an ontology of the legal domain, LKIF (Hoekstra et al., 2007), and an ontology covering the Wikipedia knowledge, YAGO (Suchanek et al., 2007). We have also explored the use of different flavors of word embeddings to transfer a Wikipedia-based model to judgments of the ECHR. We present extensive evaluation of the tools. For Argument Mining, we are manually annotating a corpus of judgments of the ECHR, with the focus on inter-annotator agreement and the performance of automatic analyzers to approach a balance between the descriptive adequacy and the performance of analyzers.

In the following Section, we outline the roadmap of our proposal, and then we go on to describe the tools and resources we are developing for NERC and NEL (Section 3.) and for Argument Mining (Section 4.), comparing them with the existing approaches in these domains. Conclusions end the paper.

2. Objectives of Information Extraction within MIREL

The final goal within the Information Extraction area of MIREL is to obtain a representation of legal texts that shows their arguments and anchors them semantically. To do that, our main subgoals are:

[1] `http://mirelproject.eu/`

- identify Named Entities and link them to a domain ontology, thus providing semantics, and
- identify argument components and their relations.

Argument Mining aims to discover the argumentative structure of a text. In the case of judgments, understanding the argumentative structure is crucial for legal actors (attorneys, judges) to make a judgment actionable in other legal actions, for example, to use the judgment as case-law. However, Argument Mining is a difficult task, even more so in the legal domain, where texts have very complex syntactical structures and semantic distinctions are very precise. Moreover, Argument Mining does not specifically deal with the propositional content of argument components. Identifying arguments does not usually include obtaining a subject-matter representation of the content of components (vs. their discursive, argumentative representation). However, for targeted applications, for higher-level analysis and for reasoning techniques we require that the propositional content is integrated with argumentative information. To do that, we build upon Information Extraction techniques.

Information Extraction is typically implemented in a pipeline. The first building block of this pipeline is usually Named Entity Recognition and Classification (NERC). The extension of NERC that anchors Named Entities to external knowledge bases, like ontologies, is known as Named Entity Linking (NEL). There are many domains where NERC achieves a good performance, and it has been shown to have a very positive impact in many applications (information retrieval, machine translation), even without any other Information Extraction technique. In particular for the legal domain, it has been shown to positively impact the identification of claims in legal texts (Surdeanu et al., 2010).

We consider the relation between NERC and Argument Mining within legal texts analogous to that of NERC and event detection in non-argumentative texts, like biomedical articles. Indeed, in both cases NERC provides an anchor to ontology-based semantics, but the relation between higher-level units is left for some other module. In factual texts, the relevant unit is the fact, which can be more or less equaled to a proposition. In contrast, in argumentative texts the relevant unit is the argument component, which can be thought of as the basic building block for applications like a reading aid, information retrieval, structured summarization.

NERC and NEL are highly domain-dependent tasks. That is why a legal NERC/NEL requires specific resources. However, developing such resources specifically for the legal domain is very costly. We have implemented a low-cost approach to legal NERC and NEL that takes advantage of the Wikipedia as an annotated corpus, more concretely, of the portion of the Wikipedia that belongs to the legal domain. To do that, we have implemented a mapping between an ontology of the legal domain, LKIF, and the YAGO ontology that is linked to the Wikipedia. This has resulted in the additional benefit of populating LKIF, which is a rather abstract ontology, and enriching its connection to Linked Open Data at more levels than the top of the ontology.

The workflow of our approach to analyze arguments in legal texts is as follows:

1. pre-process documents
2. identify and classify Named Entities
3. anchor Named Entities to a domain ontology
4. syntactico-semantic analysis of sentences, propositional representation
5. identify argument components
6. identify relations between argument components

The result of this process will be a useful input for applications like reading aids, information retrieval, structured summarization or reasoning.

In what follows we describe the tools and resources we are developing to deal with NERC and NEL and Argument Mining in the legal domain.

3. Named Entity Recognition and Linking

In this section we describe our approach to NERC and NEL.

In the legal domain, Named Entities are not only names of people, places or organizations, as in general-purpose NERC. Named Entities are also names of laws, of typified procedures and even of concepts. Named Entities may also be classified differently, for example, countries and organizations are classified as *Legal Person*, as can be seen in the following example extracted from a judgment of the European Court of Human Rights[2]:

Example 3..1 *The* $[Court]_{organization}$ *is not convinced by the reasoning of the* $[$*combined divisions of the Court of Cassation*$]_{organization}$*, because it was not indicated in the* $[judgment]_{abstraction}$ *that* $[$*Eğitim-Sen*$]_{person}$ *had carried out* $[$*illegal activities*$]_{abstraction}$ *capable of undermining the unity of the* $[$*Republic of Turkey*$]_{person}$*.*

We take an unexpensive approach to build a NERC/NEL system, by exploiting the information already available in Wikipedia as annotated examples, and connecting it with an ontology of the legal domain. More concretely, we aligned the WordNet- and Wikipedia-based YAGO ontology[3] (Suchanek et al., 2007) and the LKIF ontology[4] (Hoekstra et al., 2007) specifically conceived for representing legal knowledge. By doing this, we are transferring the semantics of LKIF to Wikipedia entities and populating the LKIF ontology with Wikipedia entities and their mentions. At the same time, we obtain a high number of manually annotated examples, taking linked strings in the Wikipedia as examples of entity mentions. With these examples, we can automatically learn a Named Entity Recognizer, Classifier and Linker.

We see that, while results on Wikipedia documents are good, there is a drop in performance when we change the domain and apply NERC to judgments of the European Court of Human Rights (ECHR). To deal with this domain change, we have explored the usage of word embeddings, without much improvement. After an analysis of error, we have identified a number of factors that will most probably impact in significant improvements.

[2] Extracted from the case Eğitim ve Bilim Emekçileri Sendikası v. Turkey, ECHR, Second Section, 25 September 2012, http://hudoc.echr.coe.int/eng.

[3] www.yago-knowledge.org/

[4] http://www.estrellaproject.org/lkif-core/

3.1. Aligning ontologies to acquire examples from the Wikipedia

The Wikipedia is a source of manually annotated examples, if we consider linked strings in the Wikipedia as examples of entity mentions. To gain access to those examples in the Wikipedia that belong to the legal domain, we aligned the WordNet- and Wikipedia-based YAGO ontology[5] (Suchanek et al., 2007) and the LKIF ontology[6] (Hoekstra et al., 2007) of the legal domain.

On the one hand, LKIF (Hoekstra et al., 2007) is an abstract ontology describing a core of basic legal concepts, with a total of 69 law-specific classes. It covers many areas of the law, but it is not populated with concrete real-world entities. On the other hand, YAGO is a knowledge base automatically extracted from Wikipedia, WordNet, and GeoNames, and linked to the DBpedia ontology[7] and to the SUMO ontology[8]. It represents knowledge of more than 10 million entities, and contains more than 120 million facts about these entities, tagged with their confidence. This information was manually evaluated to be above 95% accurate.

In our alignment process, we do not map relations but only classes. The mapping was carried out using the following methodology: for each LKIF concept, we try to find an equivalent in YAGO. If there is no direct equivalent, then we try to find a subclass, if not, a superclass. When some equivalent concept has been found, we establish the alignment using the OWL primitives `equivalentClass` and `subClassOf`. Finally, we navigate YAGO to visit the related concepts and check whether they could be aligned with another LKIF concept or if they were correctly represented as children of the selected concept. This implies that some legal concepts in YAGO are not in our ontology because they were not represented in LKIF. This is the case, for example, of the subdomain of *Procedural Law* or *Crime*, which were two annotate entities in the judgments of the ECHR. We can expect that whenever the ontology is applied to a specific subdomain of the law, it will need to be extended with the relevant concepts.

Of 69 law-specific classes within the LKIF ontology, 30 could be mapped to a YAGO node, either as children or as equivalent classes, thus 55% of the classes of LKIF could not be mapped to a YAGO node, because they were too abstract (i.e., *Normatively_Qualified*), there was no corresponding YAGO node circumscribed to the legal domain (i.e., *Mandate*), there was no specific YAGO node (i.e., *Mandatory_Precedent*), or the YAGO concept was overlapping but not roughly equivalent (as for "*agreement*" or "*liability*").

From YAGO, 47 classes were mapped to a LKIF class, with a total of 358 classes considering their children, and summing up 4'5 million mentions. However, the number of mentions per class is highly skewed, with only half of YAGO classes having any mention whatsoever in Wikipedia text.

[5] `www.yago-knowledge.org/`

[6] `http://www.estrellaproject.org/lkif-core/`

[7] `http://wiki.dbpedia.org/`

[8] `http://www.adampease.org/OP/`

The LKIF and YAGO ontologies are very different, and the task of NERC and NEL also differ from each other. In order to assess the performance of the classification at different levels, we established some orthogonal divisions in our ontology, organized hierarchically and effectively establishing different levels of granularity for the NERC and NEL algorithms to work with.

1. NER (2 classes): The coarsest distinction, it distinguishes NEs from non-NEs.
2. NERC (6 classes): Instances are classified as: Abstraction, Act, Document, Organization, Person or Non-Entity.
3. LKIF (69 classes, of which 21 have mentions in the Wikipedia): Instances are classified as belonging to an LKIF node.
4. YAGO (358 classes, of which 122 have mentions in the Wikipedia): Instances are classified as belonging to the most concrete YAGO node possible (except an URI), which can be either child of a LKIF node or an equivalent (but it is never a parent of an LKIF node).
5. URI (174,913 entities): Entity linking is the most fine-grained distinction, and it is taken care of by a different classifier, described in Section 3.3..

Example 3..1 can be tagged for NEL as follows:

Example 3..2 *The* $[\textit{Court}]_{European\_Court\_of\_Human\_Rights}$ *is not convinced by the reasoning of the* $[\textit{combined divisions of the Court of Cassation}]_{YargitayHukukGenelKurulu}$*, because it was not indicated in the* $[\textit{judgment}]_{Court\_of\_Cassation's\_judgment\_of\_22\_May\_2005}$ *that* $[\textit{Eğitim-Sen}]_{Education\_and\_Science\_Workers\_Union\_(Turkey)}$ *had carried out* $[\textit{illegal activities}]_{\emptyset}$ *capable of undermining the unity of the* $[\textit{Republic of Turkey}]_{Turkey}$.

The mapping between LKIF and YAGO is available at `https://github.com/PLN-FaMAF/legal-ontology-population`.

To build our corpus, we downloaded a XML dump of the English Wikipedia[9] from March 2016, and we processed it via the WikiExtractor (of Pisa, 2015) to remove all the XML tags and Wikipedia markdown tags, but leaving the links. We extracted all those articles that contained a link to an entity of YAGO that belongs to our mapped ontology. We considered as tagged entities the spans of text that are an anchor for a hyperlink whose URI is one of the mapped entities. We obtained a total of 4,5 million mentions, corresponding to 102,000 unique entities. Then, we extracted sentences that contained a mention of a named entity.

3.2. Learning a NERC

Using this corpus, we trained a classifier for Named Entity Recognition and Classification. The objective of this classifier is to identify in naturally occurring text mentions the Named Entities belonging to the classes of the ontology,

[9] `https://dumps.wikimedia.org/`

and classify them in the corresponding class, at different levels of granularity.
We have applied different approaches to exploit our annotated examples: a Support Vector Machine (SVM), the Stanford CRF Classifier for NERC (Stanford NLP Group, 2016), and a neural network with a single hidden layer, smaller than the input layer. We have explored more complex configurations of the neural network, including Curriculum Learning (Bengio et al., 2009), a learning strategy that is specially adequate for hierarchically structured problems like ours, with subsequent levels of granularity. However, none of these more complex configurations improved performance. For more details about the use of Curriculum Learning in our NERC, refer to (Cardellino et al., 2017).

3.3. Learning a NEL

The Named Entity Linking task consists in assigning YAGO URIs to the Wikipedia mentions. The total number of entities found in the selected documents is too big (174,913) to train a classifier directly. To overcome this problem, we use a two-step classification pipeline. Using the NERC provided by the previous step, we first classify each mention as its most specific class in our ontology. For each of these classes, we train a classifier to identify the correct YAGO URI for the instance using only the URIs belonging to the given class. Therefore, we build several classifiers, each of them trained with a reduced number of labels. Each classifier is trained using only entity mentions for a total of 48,353 classes, excluding the 'O' class.
The classifiers learnt for each of the classes were Neural Network classifiers with a single hidden layer, of size 2*number of classes with a minimum of 10 and a maximum of 500. Other classifiers, in particular, the Stanford NERC, cannot handle the high number of classes.
As a comparison ground, we also evaluated two baselines, a random classifier and a k-nearest neighbors. For the random baseline, given the LKIF class for the entity (either ground truth or assigned by an automated NERC), the final label is chosen randomly among the YAGO URIs seen for that LKIF class in the training set, weighted by their frequency. The k-nearest neighbors classifier is trained using the current, previous and following word tokens, which is equivalent to checking the overlap of the terms in the entity.
We distinguish two types of evaluations: the performance of each classifier, using ground truth ontology classes, and the performance of the complete pipeline, accumulating error from automated NERC. The individual classifier performance is not related to the other classifiers, and is affected only by the YAGO URIs in the same LKIF class. It is calculated using the test set associated with each class, that does not include the 'O' class.

3.4. Word Embeddings for Transfer Learning

The experiments were also carried out using word embeddings. Word embeddings provide a representation of words that counters the overfitting that is found in small corpora. Word embeddings are known to be particularly apt for domain transfer, because they provide some smoothing over the obtained model, preventing overfitting to the training set. Therefore, we expect them to be useful to transfer the models obtained from Wikipedia to other corpora, like the judgments of the ECHR.
However, it is also known that embeddings are more adequate the bigger the corpus they are learnt from, and if the corpus belongs to the same domain to which it will be applied. In our case, we have a very big corpus, namely Wikipedia, that does not belong to the domain to which we want to apply the embeddings, namely the judgments. Therefore, we have experimented with three kinds of embeddings: embeddings obtained from Wikipedia alone (as described above), those obtained with the same methodology but from the judgments alone, and those obtained with a mixed corpus made of judgments of the ECHR, and a similar quantity of text from Wikipedia.
The Wikipedia embeddings were obtained from the corpus we later use for the NERC task. To train word embeddings for judgments of the ECHR, we obtained all cases in English from the ECHR's official site available on November 2016, leading to a total of 10,735 documents.
All embeddings were trained using Word2Vec's skip-gram algorithm. All words with less than 5 occurrences were filtered out, leaving roughly 2.5 million unique tokens (meaning that a capitalized word is treated differently than an all lower case word), from a corpus of 1 billion raw words. The trained embeddings were of size 200, and taking them we generate a matrix where each instance is represented by the vector of the instance word surrounded by a symmetric window of 3 words at each size. Thus, the input vector of the network is of dimension 1400 as it holds the vectors of a 7 word window total.

3.5. Performance of NERC and NEL

To evaluate the performance, we computed accuracy, precision and recall in a word-to-word basis in the test portion of our Wikipedia corpus, totalling 2 million words of which the half belong to NEs and the other half to non-NEs.
For this particular problem, accuracy does not throw much light upon the performance of the classifier because the performance for the majority class, non-NE, eclipses the performance for the rest. To have a better insight on the performance, the metrics of precision and recall are more adequate. We calculated those metrics per class, and we provide a simple average without the non-NE class. Besides not being obscured by the huge non-NE class, this average is not weighted by the population of the class (thus an equivalent of macro-average). Therefore, the differences in these metrics are then showing differences in all classes, with less populated classes in equal footage with more populated ones.
Evaluating on Wikipedia has the advantage that NERC and NEL models have been learnt with Wikipedia itself, so they are working on comparable corpora. However, even if it is useful to detect NEs in the Wikipedia itself, it is far more useful for the community to detect NEs in legal corpora like norms or case-law. That is why we have manually annotated a corpus of judgments of the European Court of Human Rights, identifying NEs that belong to classes in our ontology or to comparable classes that might be added to the ontology. This annotated corpus is useful to evaluate the performance of the developed NERC and NEL tools,

but it will also be used to train specific NERC and NEL models that might be combined with Wikipedia ones.
More precisely, we annotated excerpts from 5 judgments of the ECHR, obtained from the Court website[10] and totalling 19,000 words. We identified 1,500 entities, totalling 3,650 words. Annotators followed specific guidelines, inspired in the LDC guidelines for annotation of NEs (Linguistic Data Consortium, 2014). Annotators were instructed to classify NEs at YAGO and URI levels. The annotated documents are available at `https://github.com/PLN-FaMAF/legal-ontology-population`.

3.5.1. NERC results on Wikipedia

approach	accuracy	precision	recall	F1
NER (2 classes)				
SVM	1.00	.54	.06	.11
Stanford NER	.88	.87	.87	.87
NN	1.00	**1.00**	**1.00**	1.00
NN+WE	.95	.95	.95	.95
NERC (6 classes)				
SVM	.97	.37	.18	.24
Stanford NER	.88	.78	.82	.79
NN	.99	**.89**	**.83**	**.86**
NN+WE	.94	.84	.78	.81
LKIF (21 classes)				
SVM	.93	.53	.26	.35
Stanford NER	.97	**.84**	**.71**	**.77**
NN	.97	.73	.65	.69
NN+WE	.93	.67	.60	.63
YAGO (122 classes)				
SVM	.89	.51	.25	.34
Stanford NER	–	–	–	–
NN	.95	**.76**	**.64**	**.69**
NN+WE	.90	.68	.61	.64

Table 1: Results for Named Entity Recognition and Classification on the test portion of the Wikipedia corpus.

The results for NERC on the test portion of our Wikipedia corpus at different levels of abstraction are reported in Table 1. We show the overall accuracy (taking into consideration the 'O' class), and the average recall, precision and F-measure across classes other than the non-NE class. The Stanford NERC could not deal with the number of classes in the YAGO level, so it was not evaluated in that level. We also show results with handcrafted features and with word embeddings obtained from the Wikipedia.
At bird's eye view, it can be seen that the SVM classifier performs far worse than the rest, and also that word embeddings consistently worsen the performance of the Neural Network classifier. The Stanford NERC performs worse than the Neural Network classifier at the NER level, but they perform indistinguishably at NERC level and Stanford performs better at LKIF level. However, it can be observed that the Neural Network performs better at the YAGO level than at the LKIF level, even though there are 122 classes at the YAGO level vs. 21 classes at LKIF level.

[10] `hudoc.echr.coe.int`

3.6. NERC results on the judgments of the ECHR

The results for NERC in the corpus of judgments of the ECHR are shown in Table 2. We can see the results with the models trained on Wikipedia and applied to the ECHR documents, and with models trained with and applied to the ECHR corpus (divided in training and test splits). We can also see models working on different representations of examples. The variations are handcrafted features and different combinations of embeddings: obtained from Wikipedia alone, obtained from the judgments of the ECHR alone, and obtained from Wikipedia and the ECHR in equal parts.
We can see that, on the ECHR corpus, results obtained for models trained with the annotated corpus of ECHR judgments perform significantly better than those trained with Wikipedia, even if the latter are obtained with a much bigger corpus. This drop in performance is mainly due to the fact that the variability of entities and the way they are mentioned is far smaller in the ECHR than in Wikipedia. There are fewer unique entities and some of them are repeated very often (e.g., "Court", "applicant") or in very predictable ways (e.g., cites of cases as jurisprudence).
For models trained with the annotated corpus of ECHR judgments, word embeddings decrease performance. This results are mainly explainable because of overfitting: word embeddings prevent overfitting, and are beneficial specially in the cases of very variable data or domain change, which is not the case when the NERC is trained with the ECHR corpus, with very little variability.
We also highlight that there is little difference between word embeddings trained with different inputs, although Wikipedia-trained word embeddings present better performance in general. There is no consistent difference between mixed and ECHR trained embeddings. In contrast, in Wikipedia-trained models, ECHR and mixed (ECHR+Wikipedia) word embeddings improve both precision and recall. This shows that, when we have a domain-specific model, embeddings obtained from a significantly bigger corpus are more beneficial. However, when no in-domain information is available, a representation obtained from many unlabeled examples yields a bigger improvement. For a lengthier discussion of these results, see Teruel and Cardellino (2017) (Teruel and Cardellino, 2017).

3.7. NEL results on Wikipedia

NEL could not be evaluated on the corpus of judgments, but only on Wikipedia, because annotation at the level of entities has not been consolidated in the corpus of judgments of the ECHR. Therefore, approaches to NEL have only been evaluated on the test portion of the corpus of Wikipedia.
Results are shown in Table 3. As could be expected from the results for NERC, word embeddings worsened the performance of prediction. We can see that the performance of NEL is quite acceptable if it is applied on ground-truth labels, but it only reaches a 16% F-measure if applied over automatic NERC at the YAGO level of classification. Thus, the fully automated pipeline for NEL is far from satisfactory. Nevertheless, we expect that improvements in YAGO-level classification will have a big impact on NEL.
We also plan to substitute the word-based representation of

		NERC (6 classes)				LKIF (21 classes)				YAGO (122 classes)			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Wiki trained	NN	**.76**	**.56**	**.24**	**.25**	**.76**	**.13**	**.07**	**.08**	**.76**	**.06**	**.03**	**.03**
	NN+WE wiki	.73	.34	.21	.21	.74	.08	.05	.05	.74	.03	.02	.02
	NN+WE mix	.75	.42	.23	.23	.75	.10	.06	.06	.75	.04	.04	.03
	NN+WE echr	.75	.38	.24	.24	.75	.11	.07	.07	.74	.04	.03	.03
	Stanford	.73	.36	.17	.16	.73	.07	.06	.05	-	-	-	-
ECHR trained	NN	**.80**	**.69**	.41	.47	**.81**	.46	.24	.28	**.81**	**.33**	.18	**.21**
	NN+WE echr	.77	.52	.54	.52	.75	.27	.32	.27	.79	.22	.22	.19
	NN+WE wiki	.78	.54	**.58**	.55	.79	.30	**.34**	.29	.80	.24	.22	.19
	NN+WE mix	.77	.48	.50	.48	.77	.28	.32	.28	.78	.23	**.22**	.18
	Stanford	.79	.67	.51	**.56**	**.81**	**.49**	.30	**.34**	.80	.28	.21	**.21**
	K-NN	.73	.54	.49	.50	.73	.32	.27	.25	.72	.22	.18	.16

Table 2: Results for Named Entity Recognition and Classification on the corpus of judgments of the ECHRwith models trained only with the documents of the ECHR and with models trained with the Wikipedia, combined with embeddings obtained from the Wikipedia, from the ECHR or from both.

approach	accuracy	precision	recall	F1
NEL on ground truth				
NN	.94	.48	.45	.45
NN+WE	.72	.25	.25	.25
NEL on automatic YAGO-level NERC				
NN	.69	.18	.15	.16
baselines				
Random	.51	.00	.00	.00
K-nn	.71	.14	.10	.10

Table 3: Results for Named Entity Linking on the test portion of the Wikipedia corpus.

NEs by a string-based representation that allows for better string overlap heuristics and a customized edit distance for abbreviation heuristics.

4. Argument Mining

In this section, we describe the annotation of a corpus to train Argument Mining tools. The corpus is composed of judgments of the European Court of Human Rights (ECHR) in English, obtained from the Court website[11]. This will allow us to compare our annotation to that of (Mochales Palau and Moens, 2009)[12].

We are currently working in a delimitation of the scope of annotation that provides a balance between descriptive adequacy and performance of analyzers. To approach that balance, we are analyzing inter-annotator agreement and also discrepancies between human and automated annotators, to identify concepts that produce inconsistencies and produce a more useful delimitation, in a cycle *training of annotators – annotation – analysis of discrepancies – refining of annotation guidelines*. We are currently undergoing extensive annotation of this corpus after a first iteration of this cycle.

[11] `hudoc.echr.coe.int`

[12] The dataset described in this paper is not available online.

4.1. Objectives of annotation of argumentative structure

The objective of our annotation is to identify arguments composed by claims and premises that are related to each other. Our annotation scheme is loosely based on (Toulmin, 2003), following the main adaptations that (Habernal, 2014) proposes to take the concepts from a theoretical model to practical annotation guidelines. Argument components are classified as *claims* or *premises*, with some genre-dependent attributes associated to each of these classes. The category of *major claim* is not distinguised in our annotation guidelines, as it was the main source of disagreement between annotators and it was not crucial for descriptive adequacy or application needs (Teruel et al., 2018).

The basic concepts of our annotation are:

Claim : a controversial statement whose acceptance depends on premises that support or attack it. Claims are the central components of an argument and they either support or attack the major claim. We associate each claim with the actor that has issued it.

Premise : they are the reasons given by the author for supporting or attacking the claims. They are not controversial but factual. Specifically for this corpus, We distinguish subclasses of Premises: Facts, Principles of Law and Case-law.

Argument components are connected to each other by relations, mainly *support* or *attack* relations (Simari and Rahwan, 2009). Claims support or attack other claims or a major claim, premises may support or attack claims or other premises. Additionally, we have established two more minor relations, specific for this corpus: *duplicate* (holding between claims or premises) and *citation* (holding between premises, when one cites a reference Case-law).

We have used brat (Stenetorp et al., 2012) as a tool for annotation. The guidelines for annotation, together with the annotated texts, are available at `https://github.com/PLN-FaMAF/ArgumentMiningECHR`.

4.2. Consistency of annotation, manual and automatic

For the first iteration of the cycle *training of annotators – annotation – analysis of discrepancies – refining of annotation guidelines*, four human annotators annotated 7 judgments from the ECHR, totaling 28,000 words. Approximately half of the words were annotated as belonging to an argument component.

We found a high agreement between annotators to determine whether a sentence contained an argument component, with Cohen's kappa ranging between $\kappa = .77$ and $\kappa = .84$. When this agreement is considered at token level, it varies between $\kappa = .59$ and $\kappa = .84$. We note that most disagreements occur between annotators that annotate less or more proportion of words as argumentative. Indeed, some annotators tend to consider more spans of text as argument components than others. However, there is a high agreement on spans identified as argumentative by annotators that consider less spans of text as argumentative. This has been addressed in the second version of the guidelines by a more application-oriented definition of argumentative text, focusing on an information retrieval scenario.

For the classification of argument components as premises or claims we found an agreement, ranging from $\kappa = .48$ to $\kappa = .51$ and from $\kappa = .56$ to $\kappa = .64$. We found that claims issued by the ECHR are a major source of disagreement, because the concept is mixed with that of fact or principle of law. This can be expected, as claims by a court in a judgment do have the status of principles of law after the judgment is issued, and principles of law have the same status as facts in a reasoning by a court. However, epistemologically these three concepts are difficult to reconcile. To a minor extent, claims issued by the government tend to be mixed with premises labeled as facts. Moreover, the category of premise as fact also accumulates a high number of disagreements with the category of non-argumentative text. There is also some confusion between premises interpreted as facts or as case-law, and also between premises considered case-law or law principles.

To assess the level of agreement for relations, we looked into relations that held between argument components where two annotators agreed. That meant between 46% and 74% of the components. For those, annotators agreed on the existence of a relation between components only in between 10% and 19% of the cases. When they agreed that a relation held between a given pair of components, annotators tended to agree on whether the relation was of attack, support or citation, with agreement ranging from 85% to 100% in most cases. However, the number of cases where such analysis could be carried out is so small that we require a bigger corpus to obtain more significant figures and draw conclusions upon them.

We also explored the relation between inter-annotator agreement and the performance of an automated classifier relying on the Argument classifier developed by (Eger et al., 2017), a neural end-to-end argumentation mining system with a multi-task learning setup. This system has been trained with part of the corpus, then annotated a different part of the corpus and its predictions compared with human annotations.

The comparison of human and automatic annotations is shown in Figure 1. We can see that the confusion between premises and non-argumentative text is higher than the confusion between claims and non-argumentative text, and the confusion between premises and non-argumentative text is also higher than the confusion between claims and non-argumentative text. In consequence, there seems to be a strong relation between disagreements between humans and misperformance of automatic analyzers. Addressing the first will probably have a very positive impact on the second. To address that, we have developed a refined version of the annotation guidelines, with more adequate and accurate definitions of concepts, and are currently working on annotating judgments with these guidelines.

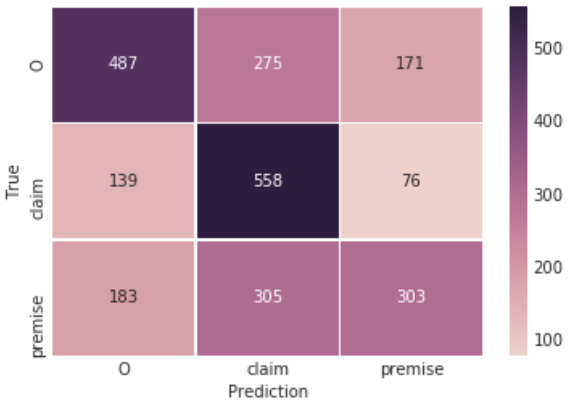


Figure 1: Confusion matrices for annotations of components between an automatic classifier and the human gold standard.

5. Summary of objectives and contributions

We have presented a work in progress for Named Entity Recognition, Classification and Linking and Argument Mining for the legal domain within the MIREL project. We have described our methodology to obtain a tool for NERC/NEL with little effort, and showed that results are promising. We have also described our approach to Argument Mining, where we are currently working on improving the annotation process to find a balance between descriptive adequacy and performance of analyzers. All tools and resources developed or in development are available at `https://github.com/PLN-FaMAF/legal-ontology-population` and `https://github.com/PLN-FaMAF/ArgumentMiningECHR`.

As future work we will improve the NERC/NEL by incorporating manually annotated examples from the ECHR, which has shown to produce good results. To optimize the annotation procedure, we will apply active learning techniques. We will also continue developing the corpus annotated for argument mining, to exploit it to train different kinds of learners, with a special focus on interpretability (i.e., Attention Networks (Cho et al., 2015)) and semi-supervised approaches (i.e., Ladder networks (Rasmus et al., 2015)).

6. Bibliographical References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 41–48, New York, NY, USA. ACM.

Cardellino, C., Teruel, M., Alemany, L. A., and Villata, S. (2017). Legal NERC with ontologies, wikipedia and curriculum learning. In Mirella Lapata, et al., editors, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers, pages 254–259. Association for Computational Linguistics.

Cho, K., Courville, A. C., and Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. CoRR, abs/1507.01053.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. CoRR, abs/1704.06104.

Habernal, I., (2014). Argumentation in User-Generated Content: Annotation Guidelines. Ubiquitous Knowledge Processing Lab (UKP Lab) Computer Science Department, Technische Universität Darmstadt, April.

Hoekstra, R., Breuker, J., Bello, M. D., and Boer, A. (2007). The lkif core ontology of basic legal concepts. In Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007).

Linguistic Data Consortium. (2014). Deft ere annotation guidelines: Entities v1.7. http://nlp.cs.rpi.edu/kbp/2014/ereentity.pdf.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. ACM Trans. Internet Techn., 16(2):10:1–10:25.

Mochales Palau, R. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009), Twelfth international conference on artificial intelligence and law (ICAIL 2009), Barcelona, Spain, 8-12 June 2009, pages 98–109. ACM.

of Pisa, M. U. (2015). Wikiextractor. http://medialab.di.unipi.it/wiki/Wikipedia_Extractor.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. (2015). Semi-Supervised Learning with Ladder Networks. July.

Guillermo Ricardo Simari et al., editors. (2009). Argumentation in Artificial Intelligence. Springer.

Stanford NLP Group. (2016). Stanford named entity recognizer (ner). http://nlp.stanford.edu/software/CRF-NER.shtml.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web, WWW '07, pages 697–706, New York, NY, USA. ACM.

Surdeanu, M., Nallapati, R., and Manning, C. D. (2010). Legal claim identification: Information extraction with hierarchically labeled data. In Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT-2010), Malta, May.

Teruel, M. and Cardellino, C. (2017). n-domain or out-domain word embeddings? a study for legal cases. In Proceedings of the ESSLLI 2017 Student Session, Toulouse, France.

Teruel, M., Cardellino, F., Cardellino, C., and Villata, S. (2018). Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18). European Language Resources Association (ELRA), may.

Toulmin, S. E. (2003). The Uses of Argument. Cambridge University Press, July.

Winick, E. (2017). Lawyer-bots are shaking up jobs. MIT Technology Review, 12. https://www.technologyreview.com/s/609556/lawyer-bots-are-shaking-up-jobs/.