D2.1

Collection of state-of-the-art NLP tools for processing of legal text

Grant Agreement nº:	690974
Project Acronym:	MIREL
Project Title:	MIning and REasoning with Legal texts
Website:	http://www.mirelproject.eu/
Contractual delivery date:	31/12/2016
Actual delivery date:	18/01/2017
Contributing WP	WP2
Dissemination level:	Public
Deliverable leader:	UL
Contributors:	UL, UNITO, UNIBO, INRIA, DATA61

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690974

Document History

Version	Date	Author	Partner	Description
0.1	15/12/2016	Livio Robaldo	UL	Initial draft
0.2	27/12/2016	Livio Robaldo	UL	Second draft
0.3	10/1/2017	Livio Robaldo	UL	Third draft
1.0	18/1/2017	Livio Robaldo	UL	Final Version

Contributors

Partner	Name	Role	Contribution
UL	Livio Robaldo	[Editor/Coordinator]	Literature review, coordinator of the drafts
UNITO	Luigi Di Caro	Reviewer	Literature review, sections contributor
UNIBO	Monica Palmirani	Reviewer	Literature review, sections contributor
INRIA	Serena Villata	Reviewer	Literature review, sections contributor
DATA61	Guido Governatori	Reviewer	Literature review, sections contributor

Table of Contents

Executive Summary

The structure of this deliverable is rather simple. After an introduction offering a general overview of available NLP technologies, the deliverable lists and briefly describe the main ones, together with examples of projects, applications, or research activities in legal informatics where they are used.

1 State-of-the-art NLP tools

NLP technologies are indispensable components of applications in legal informatics, as legal knowledge is originally available in natural language only. MIREL partners are not concerned with the availability of the Language Technology per se, but with the adaptation of the existing tools and resources for the legal domain.

Current approaches in NLP have made explicit that the processing heavily depends on the domains. In-domain solutions are more stable than out-of-domain ones. Legal texts in particular use specific lexicon and natural language expressions that rarely occur in non-legal texts. Available NLP tools are, by and large, language-dependent: while there are basic NLP tools (pos-tagger, parsers, etc.) with good performances for English, in that it is the most worldwide used language, for other languages, e.g. Italian, even if they are rather widespread, the availability is limited.

NLP techniques may be classified in two main classes:

- **Statistical techniques**, used to look for *general information* in the documents. For instance, for associating a document, an article, a paragraph, etc. with certain labels (from a finite set of available label) that describes the topic it is about (classification).
- **Rule-based techniques**, used to look for *specific information* in the documents. For instance, for extracting references to legal documents, dates, locations, names of persons, etc., as well as certain relevant concepts (entity linking).

This is an approximation of the actual state of the affairs, which is only useful here to define the borders of the state of the art as well as the scope of the MIREL project.

In the literature there are rule-based technique used to classify or, more generally, to look for general information. An example is [Robaldo et al, 2012], where a rule-based procedure has been proposed to classify modificatory provisions in legal texts, e.g. abrogation, substitution, etc.

It is also true that statistical techniques are also adopted in case of specific information extraction scenarios, and they are based on the building of statistically-significant patterns that aim at covering the expression of specific lexical entities. For example, the work in [Boella et al., 2014] is of this type.

Finally, of course it is possible to devise hybrid approaches, i.e. statistical approaches that output set of rules, which can be manually tuned afterwards. This direction of research is currently under investigation in the TULE parser [Lesmo, 2007].

It is well-known that statistical and rule-based techniques are in trade-off.

Often, statistical NLP tools need initial corpus of annotated data on which it is possible to train models that the NLP tool will use to annotate new text. Provided such corpora are available, training a statistical model to make it able to process new text is a quick phase. However, it is often the case that statistical models, once trained on big corpora, are able to properly parse new linguistic constructions but, at the same time, are no longer able to properly parse linguistic constructions that were previously handled. As a result, the performances of these parsers tend to oscillate around a certain maximum value of accuracy (usually, 80-90% of accuracy, depending on domains, volumes of data, technical parameters, and algorithms).

In other scenarios, statistical approaches may be unsupervised. In these cases, the mathematical models are aimed at extracting patterns which reflect some specific and context-based statistical significance. These patterns may include linguistic features as well as numeric frequencies, and they can be used to match (even partially) with new input data. This mechanism permits to structure unstructured data in automatic ways, in the form of clusters that share levels of internal similarity.

On the other hand, by using rules, when new linguistic constructions are found, it is sufficient to add rules to cover them. And, by evaluating the rule-based system with respect to a corpora, it is in principle possible to achieve 100% of accuracy, in that, once a new rule is added, it is easy to automatically identify which linguistic constructions are no longer handled and so how the rule has to be modified accordingly. The main disadvantage of rule-based tools is that building a suitable set of rules is a very slow and time-consuming process, because they are manually inserted in most cases.

As said above, basic NLP tools are generally multi-purpose, i.e. defined to process natural language text *in general*. In other words, available NLP tools are not developed for processing legal text. Of course, in order to have good performances, NLP tools need to be trained and tuned on legal texts, in order to capture the particularities of the legal language.

NLP tools are generally classified as following, although it is possible to find tools that integrate two or more of them.

- **Part of Speech (POS) taggers**: NLP tools that, taken a free text in input, classify each word according to its part of speech. Available POS taggers distinguish between content parts of speech (nouns, verbs, adjectives, adverbs) and functional part of speech (articles, prepositions, punctuations, etc.). Besides the part of speech, many POS taggers associate information about the inflections of the words (gender, number, mood, etc.).
- **Named Entity Recognition (NER)**: the task of seeking and classifying named entities in texts, according to a pre-defined set of categories. In most cases, these are categories of proper nouns, subcategorized in proper nouns of persons, cities, named artifacts, etc. Besides proper nouns, NER tools are able to recognize time expressions, quantities, monetary values, etc. as well as particular concepts or linguistic patterns.
- **Syntactic parsers**: NLP tools that taken the result of a POS tagger on a free text provide a description of how the words are organized within a sentence. There are two approaches:

(1) constituency approach, i.e. grouping contiguous words into phrases and (2) dependency approach, i.e. linking the words between them according to a predefined set of grammatical categories (subject, object, etc.).

- **Text Segmentation**: The goal of Text Segmentation (TS) is to identify boundaries of topic shift in a document. Discourse structure studies have shown that a document is usually a mixture of topics and sub-topics. A shift in topics could be noticed with changes in patterns of vocabulary usage [Michael Alexander et al, 2014]. The process of dividing text into portions of different topical themes is called Text Segmentation. The text units (sentences or paragraphs) making up a segment have to be coherent, i.e., exhibiting strong grammatical, lexical and semantic cohesion. Applications of TS includes Information Retrieval (IR), passage retrieval and document summarization.

- **Topic Modeling**: Topic models are fundamental tools for the extraction of regularities and patterns providing automatic ways to organise, search and give sense to large data collections. The shared basic assumption is that documents have a latent semantic structure that can be inferred from word-document distributions. Latent Semantic Analysis (LSA) [Deerwester et al., 1990] is a linear algebra-based method that reduces the a word-document co-occurrences matrix into a reduced space such that words which are close in the new space are similar. Its probabilistic and generative version (pLSA) [Hofmann, 1999] adds a latent context variable to each word occurrence which explicitly accounts for polysemy. Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is a fully Bayesian probabilistic version of LSA. Given a corpus of documents, the idea underlying LDA is that all documents share the same set of topics, but each document exhibits those topics in different proportions depending on words which are present in that document. Topics, in turn, are defined as different probability distributions over the words of a fixed vocabulary, but they are interpreted by restricting attention to words with the highest estimated frequency. Only documents are observed, while the topics, per-document topic distributions and the per-document per-word topic assignments are latent structures inferred from the data.

- **Keyphrase Extraction**. The imperative of analysing texts and extracting salient cues that describe the overall content and semantics of the text is an appeal that has found numerous applications in Natural Language Processing. Keyphrases, sometimes called keywords, are concepts that provide concise descriptions of a document content and have shown to be useful in document indexing e.g., by librarians in categorising books, document clustering and visualisation, text summarisation, text categorisation, text annotation and topic indexing and emerging topic detection.

- **Semantic annotation**: Texts, and also legal texts in this project domain, can be associated with semantic information or tags which enable semantic access, retrieval, management, and analysis of big data collections. These tags can be even structured in some way, or they can reflect an ontological organisation of concepts. Still, the association with tags can happen at different levels of granularity. In the light of this, we define two specifications:

    - Semantic Roles - when short chunks of texts play a semantic role with respect to the described situation. This seems to overlap with named entities, but even named

entities can play different roles in different sentences, so the semantic grain in this case is much deeper.

- Semantic Categories (or Classes) - this is usually the label to describe the association of medium-long texts with tags representing topics or types of documents. Semantic categories often require Machine Learning to build models able to automatically classify input texts in one or more categories based on what learnt from labeled examples (training set).

- **Question Answering**: In the restricted domain of the MIREL project, Legal Question Answering (LQA) is the task of answering to questions in natural language with respect to legal text. In particular, we mention the specific task of answering solution answers of the type "YES or 'NO', where 'YES' means that the question is entailed by a text (and 'NO' otherwise).

1.1 The Stanford parser

The Stanford NLP library[1] is one of the most used NLP tools. It provides a set of natural language tools that cover most of the techniques explained in the previous overview section. It gives the base forms of words, their parts of speech, the named entities, and normalises tokens representing dates, times, and numeric quantities. Then, it finds how terms of phrases are linked through syntactic dependencies.

The Stanford NLP framework is based on a statistical model and it is known to be reliable and fast even on large input data. It supports several languages other than English, and it can be run as a simple web service. It is one of the most used parsers in the world.
In MIREL, this library will be mostly used for the English language, while for other languages, e.g. Italian and Bulgarian, more language-specific tools are used, in order to achieve high-quality results.

1.1.1 Automatic Annotation of Semantic Entities and Roles

The partner UNITO already faced the problem of automatically extracting structured knowledge to improve semantic search and ontology creation on textual databases. UNITO has proposed and implemented an approach that first relies on the Stanford Parser to transform POS tags and syntactic dependences into generalised features that aim at capturing the surrounding linguistic variability of the target semantic units.

These new featured data are fed into a Support Vector Machine classifier that computes a model to automate the semantic annotation. The system is able to extract semantic entities and roles within general legal texts [Boella et al, 2014].

[1] http://nlp.stanford.edu/software/lex-parser.shtml

1.1.2 Finding and classifying Named Entities in legal texts

The partner INRIA together with the partner Cordoba exploited part of the Stanford NLP library for their work about ontology population of legal ontologies. The idea is to exploit Wikipedia as a source of manually annotated examples of legal entities. They align YAGO, a Wikipedia-based ontology, and LKIF, an ontology specifically designed for the legal domain. Through this alignment, they can effectively populate the LKIF ontology, with the aim to obtain examples to train a Named Entity Recognizer (NER) and Classifier to be used for finding and classifying entities in legal texts.

They evaluated their approach by training a linear classifier, namely a Support Vector Machine (SVM) with a linear kernel, and the Stanford CRFClassifier model for NER. As an additional baseline for the evaluation, they obtained the performance of the Stanford NER system, training it with their corpus with Wikipedia annotations for the LKIF classes.

1.1.3 [Huang, 2014]

The number of applying documents for global patent has been growing fast in the fastest in the last years. This system aims at categorising patent documents making use of the Stanford Parser, together with the Rough Set Theory. The reason behind the system is that, for patent document classification, manually review is not only slow but also costly. This research aims to have an efficient approach for automatic patent document classification.

[Huang, 2014] used the WIPO-alpha database from World Intellectual Property Organization for two experiments, the first one combining Stanford Parser and Rough Set Theory to build a N-gram noun phrase extraction algorithm. The second experiment aimed at combining Stop Words and Rough Set Theory. Results demonstrated that the average F-score for the proposed method was 98.5%, higher than traditional methods.

1.1.4 [Wyner et al, 2011]

In this work, the authors faced the problem of identifying, extracting, and formalising conditional and normative rules from legal regulations. For instance, there are XMLs for legal materials making them available, searchable, and linkable on the Internet such as CEN MetaLex along with national standards, for example, in the United States.

However, identifying, extracting, and formalising the rules remains a highly knowledge and labour intensive task, creating a significant bottleneck between the semantic content of the source material, expressed in natural language, and computer-based, auto- matic use of that content. To address the bottleneck, natural language processing (NLP) techniques have been applied.

The main contribution in [Wyner et al, 2011] is the identification and extraction of high level components of rules from regulations in English, applying and extending widely available, current NLP tools such as Stanford Parser. The authors then provided an open source, modular framework along with an explicit methodology and open source materials.

1.2 Apache OpenNLP

The Apache OpenNLP[2] library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy and perceptron based machine learning. Models trained on manually annotated text corpora are available in different languages.

1.2.1 Intelligent Amplification for Cybercrime (IAC)

The project Intelligence Amplification for Cybercrime (IAC) [Testerink-Bex, 2016] is a project carried out at the University of Utrecht (the Netherlands) aiming at improving the online intake of criminal complaints and the subsequent investigations on the topics of e-crime and cybercrime for the Dutch National Police. A dialogue module involving NLP technologies allows for mixed-initiative dialogues between human complainants and software agents for crime intake. OpenNLP is used for named entity recognition in the dialogue module, specifically to classify words such as personal name and organisation names. The recognized named entities are then associated with ontological concepts and relations. Further details are available in [Bex et al., 2016].

1.2.2 [Vico-Calegari, 2015]

This paper presents a software architecture for supporting document anonymization, i.e. for replacing sensitive data in such a way that preserves the confidentiality of the documents without altering their value or usefulness. OpenNLP is integrated within a system called MultiNER adapter in order to recognize named entities to anonymize, e.g. persons or geographical locations, in the jurisprudence documents published by the Courts of Appeals and the Supreme Court of Uruguay.

1.2.3 [Schilder, 2005]

This paper presents a prototype for identifying relevant events in the United States Code on U.S. immigration nationality and associating them with the temporal information, e.g. *he entered the United States before December 31, 2005*, a task that has been deemed to be pivotal for automatically processing legal documents, which are usually plenty of temporal information (cf. [Vila-Yoshino, 1998]). In the prototype, openNLP is used for carrying out POS tagging, sentence splitting, and shallow parsing on the input text before an ad-hoc rule-based system on the parsed text recognizes events, temporal constraints, and associate the former with the latter.

[2] http://opennlp.apache.org/

1.3 Spacy and TensorFlow

SpaCy[3] is a statistical tool for large-scale information extraction tasks. It is well-known for its speed in parsing very large textual input. SpaCy is designed also for real and industrial scenarios and its API is simple and productive. Its long list of features includes syntax-driven sentence segmentation, Part-of-speech tagging, Named entity recognition, easy deep learning integration, statistical models for English and German, and state-of-the-art speed. In details, Spacy has an accuracy of 92.8, while the state of the art has an accuracy of 94.44 (Andor et al, 2016).

SpaCy is also easy to integrate with TensorFlow[4], an open source software library for numerical computation using data flow graphs. In TensorFlow, nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them.

1.3.1 Alignment and Reconstruction of EuroVoc with Legal Taxonomies

The partner UNITO implemented a Memory Network-based Question Answering (QA) system which test a Machine's understanding of legal text and identifies whether an answer to a question is correct or wrong, given some background knowledge. UNITO also prepared a corpus of real bar exams for this task [Adebayo et al., 2016].

QA follows the Human learning process, i.e., committing to memory and generalizing on new events. The authors in [Sukhbaatar et al, 2015] using Deep Neural Networks achieved 100% accuracy on some tasks. However, a synthetic dataset was used and the evaluations tested the ability of the models in providing factoid answers to questions of where, when and who about an entity. The implemented system answers to the following question: "Can we use deep learning techniques to achieve transfer-learning on passage- question-answer (PQA) with similar case templates?". By transfer learning, we here mean a generalisation procedure whereby the proposed model is able to transfer hidden facts from a scenario to similar scenarios.

Deep Networks can autonomously learn semantic representation from text. Recurrent Neural Networks (RNNs) [Medsker et al, 2001] have connections that have loops, adding feedback and memory to the networks over time. However, RNNs memory are small and also not compartmentalized enough for long range information retention. [Weston et al 2014] proposed the MemN as a solution, which is employed in this system. MenNs are composed of 4 units, i.e., input units I, the Generalisation Unit G, output unit O and the response unit R, which generates a representation of the Output in any specified format. The Long Short-Term Memory (LSTM) [Hochreiter et al, 1997] is a special kind of RNNs that is robust to the vanishing gradient problem.

1.3.2 [Schrading, 2015]

The work of [Schrading, 2015] aimed at analyzing domestic abuse using SpaCy on Social Media Data. Social media and social networking play a major role nowadays. Publicly available posts on websites such as Twitter, Reddit, Tumblr, and Facebook can contain deeply personal accounts of the lives of

[3] https://spacy.io
[4] https://www.tensorflow.org/

users. Health woes, family concerns, accounts of bullying, and any number of other issues that people face every day are detailed on a massive scale online. NLP and Machine Learning may help the process of analysis to understand societal and public health issues. This allows fas data collection and analysis that can shed light on sociologically important problems.

The work of the [Schrading, 2015] work studied the efficacy of classifiers that detect text discussing abuse to highlight the dynamics of abusive relationships. Analysis revealed micro-narratives in reasons for staying in versus leaving abusive relationships. This type of method can be also adopted in the legal context.

1.4 Gensim

Gensim: Gensim is a scalability, robust and platform indipendent python library to realize unsupervised semantic modelling from plain text, such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA)[5]. Gensim can be integrated with other python libraries like NLTK (Natural Language Toolkit) for carrying out NLP pre-processing tasks. Gensim provides implementations of tf-idf (a common weighing factor in text mining), word2vec (shallow neural network models that produce word embeddings) and topic modeling algorithms.

1.4.1 Automated Transposition Detection of European Union Directives

Within the MIREL project we implemented a system based on semantic text similarity techniques to automatically detect the transposition of European Union (EU) directives into the national law. Currently, the European Commission (EC) resorts to time consuming and expensive manual methods like conformity checking studies and legal analysis for identifying national transposition measures. We utilise both lexical and semantic similarity techniques and supplement them with knowledge from EuroVoc to identify transpositions. Such systems could be used to identify the transposed provisions by both EC and legal professionals.

UNITO has developed systems for semantic similarity for automatically identifying the transposition of EU directives into the national law while collaborating with APIS. UNITO developed the semantic similarity system while APIS provides resources like directive documents in XML format and useful thesaurus and dictionaries which are integrated into the system. APIS is also developing a platform for evaluating and visualizing the results produced by the Directive transposition system. The current system utilizes Latent Semantic Analysis and knowledge from EuroVoc thesaurus to identify transpositions. The retrieved transpositions are compared with a gold standard or evaluated by APIS legal experts to compute the accuracy of the system. The system would serve as legal Information retrieval tool to support the lawyers and EC professionals carrying out cross-border legal research.

[5] https://radimrehurek.com/gensim/

The effective transposition of European Union (EU) directives at the national level is important to achieve the objectives of the Treaties and smooth functioning of the EU. Member States are responsible for the correct and timely implementation of directives. The European Commission (EC) is responsible for monitoring the national implementations to ensure their compliance with EU law. The transposition measures adopted by Member States in national legislation to achieve the objectives of the directive are known as national implementing measures (NIMs)[6]. The Commission monitors the NIMs (communicated by the Member States) to ensure that Member States have taken appropriate measures to achieve the objectives of the directive. The steps taken by the Commission to monitor NIMs include Conformity Checking and Correlation tables [Eliantonio, 2013]. The Commission outsources the monitoring of NIMs to subcontractors and legal consulting firms. The conformity check studies carried out by a team of competent legal experts, comprise legal analysis and concordance tables. The concordance tables identify the specific provisions of NIMs which implement a particular article of the directive. Correlation tables are prepared by the Member States to ensure that the directive is completely transposed. They identify the specific provisions of NIMs for each article of a directive in a tabular format. Correlation tables are generally not available to public as they are sent by Member States to the Commission as part of a confidential bilateral exchange. There is no agreed format or compulsory content for correlation tables.

The system implemented by UNITO represents the first work in automated transposition detection of EU directives. The objective is to identify the specific provisions of NIMs which transpose a particular article of the directive. UNITO compared the results from both lexical and semantic similarity techniques on five directives and their corresponding NIMs by evaluating them with a gold standard (correlation tables) with good performance.

In detail, the system uses cosine similarity vector model (lexical similarity technique) to detect transposing provisions with similar words. Latent semantic analysis (semantic similarity technique) was chosen to detect transposing provisions with same semantics but different wordings. Preprocessing included removing punctuation, conversion to lowercase and tokenization. Further stop-words were removed and part-of-speech tagger (POS tagger) was used to filter out nouns, verbs and adjectives from the remaining set of tokens. The tokens obtained after pre-processing were enriched with the knowledge from EuroVoc, a multilingual thesaurus of the European Union. The tokens in the corpus were enriched with synonym and near-synonym terms as per equivalence relationship of EuroVoc [Paredes et al, 2008]. Afterwards, the set of new tokens are stemmed to reduce the inflectional forms of words. Each provision of the corpus is then represented in a bag-of-words format. It is a list of each token and its count in a particular provision. Further, the system employs Term Frequency-Inverse Document Frequency (tf-idf) weighting scheme to all the provisions. It has been implemented a latent semantic analysis (LSA) by applying Singular Value Decomposition (SVD) to the tf-idf provision-token matrix. SVD decomposes the tf-idf matrix into

[6] European Commission. Monitoring the application of Union law, 2014 Annual Report

separate matrices which capture the similarity between tokens and provisions across different dimensions in space [Dumais, 2004]. Figure 1 shows the system architecture.

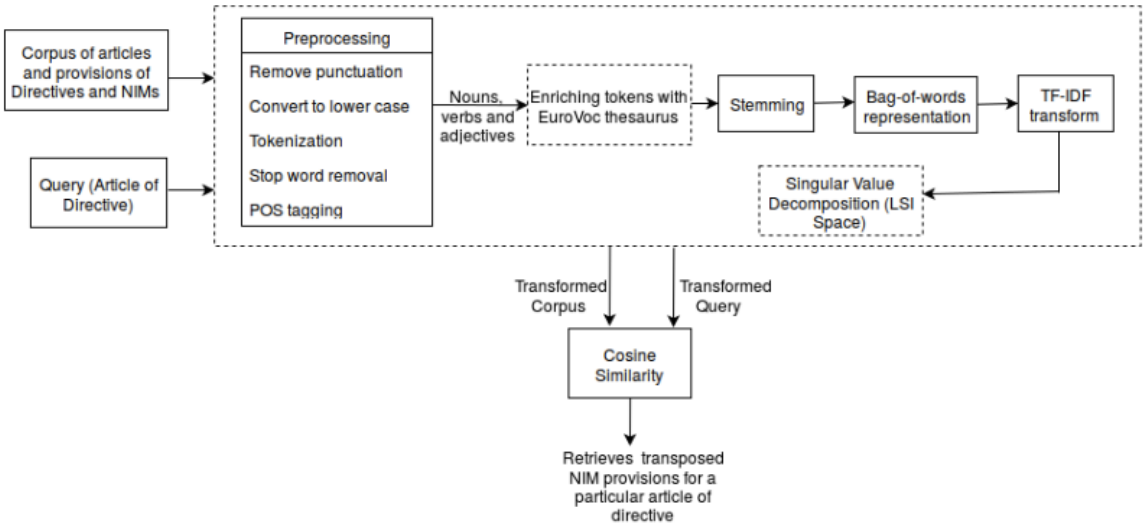

Figure 1. System architecture for automated transposition detection

1.5 Boxer/CGG parser

Boxer/CGG parser [Bos, 2008] is a syntactic-semantic NLP tool that includes a combinatory categorial grammar (CCG) parser [Steedman-Baldridge, 2011] and syntax-semantic interface from the CGG structures to semantic representations in Discourse Representation Structures (DRSs) [Kamp-Reyle, 1993].

A CCG grammar is a constituency grammar that describe syntactic typed categories of lexical items along with their mode of combination. For instance, (many) verbs are associated with the syntactic category NP/S, specifying that the verb can combine with a noun phrase (NP) in order to build a sentence (S). In Boxer, the syntactic parser is associated 1:1 with a formal semantic translation expressed in lambda-calculus such that the semantic derivations follow the structure of the syntactic tree. Those are translated in DRT, whose semantic representations are basically FOL expressions able to represent discourses, including pronominal anaphora and discourse relations. Syntactic and semantic derivations can be given for long and complex sentences as well as for discourses, which are sequences of sentences.

1.5.1 Extracting deontic rules from technical documents

Despite the numerous approaches tackling the problem of moving from a natural language legal text to the respective set of machine-readable conditions, results are still unsatisfiable and it

remains a major open challenge. The INRIA partner and the Data61 partner, in collaboration with FBK-Trento (Italy), proposed an approach to combine different Natural Language Processing techniques towards the extraction of rules from legal documents. More precisely, they combine the linguistic information provided by WordNet together with a syntax-based extraction of rules from legal texts, and a logic-based extraction of dependencies between chunks of such texts. Such a combined approach leads to a powerful solution towards the extraction of machine-readable rules from legal documents. The proposed approach is then evaluated over the Australian "Telecommunications consumer protections code".
In this work, the Stanford NLP library is used for parsing natural language sentences to retrieve their grammatical representation, and a Combinatory Categorial Grammar (CCG) parser tool including the Boxer framework is used for extracting logical dependencies between chunks of text from the document. More details about this approach are available in [Dragoni et al., 2016].

1.5.2 [Wyner et al., 2012]

In this work, Boxer/CGG has been used to parse the British Nationality Act 1981 (PART I), in order to create a corpus of legal texts. The corpus has been used to identify several linguistic phenomena occurring in legal texts, in order to advocate further research on NLP applied to the legal domain. Examples of such linguistic phenomena are generalized quantifiers, intensionality, and genericity. Furthermore, it has been identified that many cases of complex occurring in legal texts are not currently resolved by Boxer so that a further augmentation of the tool to this end is needed.

1.6 JFLEX

JFlex is a lexical analyzer generator (also known as scanner generator) for Java, written in Java. It takes as input a specification with a set of regular expressions and corresponding actions. It generates a program (a lexer) that reads input, matches the input against the regular expressions in the spec file, and runs the corresponding action if a regular expression matched. JFlex lexers are based on deterministic finite automata (DFAs). They are fast, without expensive backtracking.

1.6.1 BO-ECLI

JFLEX is being used for identification of national and european references within legal texts in the project BO-ECLI. Building on ECLI (BO-ECLI[7]) is a project involving sixteen partners from ten Member States (Italy, Greece, Croatia, Estonia, Belgium, the Netherlands, Germany, the Czech Republic, Spain, Romania) that aims to broaden the use of ECLI and to further improve the accessibility of case law. In the project, UNITO and UNIBO are supporting the creation of a framework for the identification and construction of ECLI identifiers.

[7] http://bo-ecli.eu

1.7 Parse-IT

Parse-IT is a component of the Digital Legislation Platform[8], hosted by the MIREL partner Data61-CSIRO. Parse-IT is a proprietary web-based system able to automatically analyze and translated legislation into Defeasible Deontic Logic [Governatori et al., 2013]. The tool takes as input a normative text and identifies the logical structure of the norms, and represents them in a set of rules. Parse-IT also includes a specialized rule editor to enhance, correct and fine-tune the rules extracted from legal documents. The rules can then be used perform various advanced legal reasoning tasks such as business process compliance [Governatori et al., 2009] and analysis of contracts [Governatori, 2005].

1.7.1 PermitME

Parse-IT has been used in the PermitME project, in collaboration with the Australian Taxation Office, aiming at the creation of a digital concierge to guide SMEs though the maze of legislation that applies to their business to determine what permits and licenses are required for the business and whether the requirements to apply for such permits and licenses are satisfied.

1.7.2 Regorous

Parse-IT has been integrated in the Data61 business process compliance suite Regorous[9]. The Digital Legislation Platform is meant to create a framework for the digitalization and formalization of legislation to be released as open data by government, and the creation of APIs for retrieving and reasoning with the formalized legislation for the creation of legal applications.

1.8 TULE parser

The TULE (Turin University Linguistic Environment) parser [Lesmo, 2007] has been developed at the Department of Computer Science of the University of Turin, and it is massively used by the partners UNITO and UL in MIREL to analyze legal text. The reason is that, contraty to most other existing parsers, e.g. the ones listed above, TULE is rule-based, so that it is possible to manually correct the NLP analysis without training the parser on new corpora. TULE includes a Tokenizer, a Sentence splitter, a POS Tagger, and a Dependency Parser. TULE works for Italian, English, and also other languages that are not relevant for MIREL. For Italian it has been trained on the TUT corpus, which is the biggest syntactic corpus (treebank), freely available for Italian: it contains more than 3500 sentences annotated via TULE dependency format. TUT sentences have been mainly taken from newspaper, Italian civil code, Wikipedia entries. For French and English (and also Italian), Tule has

[8] https://digitallegislation.net
[9] https://www.regorous.com

been trained on legal texts taken from JRC-Acquis[10], UDHR: (Universal Declaration of Human Rights), European Parliament Parallel Corpus [Koehn, 2005], and others.

1.8.1 ProLeMAS

The project ProLeMAS[11] (PROcessing LEgal language in normative Multi-Agent Systems) is an individual project carried out by Livio Robaldo at the university of Luxembourg. The project is supported by a Marie Skłodowska-Curie Individual fellowship. One of the aim of the project is to implement a suite of NLP tools in order to perform named entity recognition via dependency parsing, as well as for recognizing obligations, rights, etc. occurring in EU directives. TULE is the dependency parser used in the project.

1.8.2 DAPRECO

The project DAPRECO (DAta PROtection REgulation COmpliance) is a CORE project[12] that has been retained for funding on Nov 2016. The project will start on Feb 2017 and it will last two years and an half. The University of Bologna is an external partner of the project. The NLP tools developed in ProLeMAS, as well as the TULE parser, will be interfaced to the LIME[13] annotator, developed at the University of Bologna, in order to build a semi-automatic approach for annotating legal documents. See [Bartolini et al., 2016].

1.8.3 Eunomos

Eunomos is an advanced legal document management system based on legislative XML representations of laws which are retrieved automatically from institutional legislative portals. The system is described in [Boella et al, 2016]. The TULE parser is used in Eunomos in order to recognize concepts and named entities as well as for classifying the legal documents stored in the system. A commercial version of Eunomos, called MenslegiS, is patented and distributed by Nomotika SRL.

1.8.4 OpenSentenze

The project OpenSentence[14] aims at publishing case law from Italian courts as open data. In order to do so, case law need to be anonymized: the names of the parties as well as any other information allowing to identify the parties (addresses, dates of birth, etc.) must be hidden. Note that, on the other hand, proper nouns of judges and lawyers can appear in the document. The TULE parser is used in a semi-automatic way together with the LIME annotator developed at the University of Bologna. The TULE parser generates XML files in Akoma Ntoso[15] where named entities to be anonymized are suggested. A human annotator validates the suggestions or add missing annotations.

[10] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[11] http://www.liviorobaldo.com/prolemas.html
[12] https://www.fnr.lu/funding-instruments/core
[13] http://lime.cirsfid.unibo.it/
[14] http://www.opensentenze.it/
[15] http://www.akomantoso.org/

1.9 SPeLT

CIRSFID (UNIBO) has developed SPeLT (Semantic Parser of Legal Text), a framework of tools for parsing and analyzing legal texts. Currently, SPeLT supplies two main tools for parsing; SPeLT-ref and SPELT-struc. The aim of SPELT-ref is to identify legal references in judgments and other legal documents, while the aim of SPELT-struc is to identify the logical structure of legal documents. Also, SPELT supplies a set of rules, called SPELT-anony, that can be used by third-parties software in order to anonymize specific entities in judgements and other legal documents. Lastly, SPELT supplies a markup tool, called SPeLT-mark, that can use the results of a parser to markup an unstructured document with AkomaNtoso. To accomplish this, SPELT-mark can use results returned by SPELT-struc and SPELT-ref, but it can also use results returned by third-parties software.

1.9.1 EuCases

The project EuCases[16] was an FP7 project involving the MIREL partners APIS, UNITO, Nomotika, and UNIBO. Spelt NLP tools was used inside of LIME editor[17] for marking up Italian acts in order to demonstrate the effectiveness of the Akoma Ntoso for representing EU directives published in the Official Gazette. Spelt tool detects from plain text the main legal knowledge information inside of the EU legal acts as: structure of the document, number of the document, type of document, authority that emits the act, normative references, dates, persons, organizations, locations, roles.

1.9.2 Swiss Chancellery project

The project was part of an agreement between UNIBO and the Swiss Chancellery. Spelt NLP tools was used inside of LIME editor[18] for marking up Federal Chancellery of Switzerland bills, acts, referenda, consolidated code, in order to demonstrate the effectiveness of the Akoma Ntoso for managing the publication workflow in multilingual context (French, German, Italian). Spelt tool using Word stylesheets used by the end-users, can detect the main legal knowledge information inside of the Spanish documents as: structure of the document, number of the document, type of document, authority that emits the act, normative references, dates, persons, organizations, locations, roles.

1.9.3 High court of Cassation project

The project was part of an agreement between UNIBO and the Italian Court of Cassation. Spelt NLP tools was used inside of LIME editor[19] for marking up Italian acts in order to demonstrate the effectiveness of the Akoma Ntoso for representing Italian legal documents published in the Official

[16] http://eucases.eu
[17] http://lime.cirsfid.unibo.it
[18] See: http://sinatra.cirsfid.unibo.it/lime-dev-ch.
[19] See: http://sinatra.cirsfid.unibo.it/lime-cassazione/

Gazette. Spelt tool detects from plain text the main legal knowledge information inside of the Italian legal acts as: structure of the document, number of the document, type of document, authority that emits the act, normative references, dates, persons, organizations, locations, roles.

1.9.4 FAO project

The project was part of an agreement between UNIBO and the Food and Agriculture Organization of the United Nations. Spelt NLP tools was used inside of LIME editor[20] for marking up FAO documents (standards, basic texts, resolution) in order to demonstrate the effectiveness of the Akoma Ntoso for modelling the workflow of the documents and their publications. Spelt tool detects from the plain text the main documentary information inside of the legal acts as: structure of the document, number of the document, type of document, authority that emits the act, dates, persons, organizations, locations, roles, keywords.

2 Conclusions

This deliverable have reported main NLP tools used by the community in legal informatics for extracting relevant information from legal texts via various tasks. Main tasks, performed by the tools listed above, are part-of-speech tagging, named entity recognition, parsing, text segmentation, topic modeling, keyphrase extraction, semantic annotation and question answering. The report listed and briefly described main NLP tools, most of which are indeed multi-purpose tools, i.e. not specifically used for processing legal texts only, as well as specific projects and research activities in legal informatics where they have been used. The mostly used tool is perhaps the Stanford Parser, a statistical tool that performs pos-tagging and parsing on texts written in natural language, as well as named entity recognition. Other statistical NLP tools are Apache OpenNLP, Spacy, TensorFlow, and Gensim. Besides statistical tools, the community in legal informatics also adopted rule-based tools. As explained in the introduction, rule-based tools tend to have better performances than statistical one for tasks where it is pivotal to look for specific information. In legal informatics, one of the main task needing a rule-based system is perhaps the one of looking for references of legal documents within other legal documents, which may be seen as a specific task of named entity recognition. The project BO-ECLI listed above is centered on this task. Examples of rule-based NLP tools, described in this deliverable, are Boxer, JFLEX, Parse-IT, TULE parser and SPeLT.

[20]See http://sinatra.cirsfid.unibo.it/node/portalfaoresolution/ and http://sinatra.cirsfid.unibo.it/node/portalfaobasic/

References

1. [Adebayo et al., 2016] Neural Reasoning For Legal Text Understanding (Guido Boella Adebayo Kolawole John, Luigi Di Caro), In Proceedings of the 29th International Conference on Legal Knowledge and Information Systems (JURIX2016), 2016.
2. [Andor et al, 2016] Andor, Daniel, et al. "Globally normalized transition-based neural networks." arXiv preprint arXiv:1603.06042 (2016).
3. [Bartolini et al., 2016] C. Bartolini, G. Lenzini and L. Robaldo: *Towards legal compliance by correlating Standards and Laws with a semi-automated methodology*, in proc. of the 28th Annual Benelux Conference on Artificial Intelligence. Amsterdam, 2016.
4. [Bex et al., 2016] Floris Bex, Joeri Peters and Bas Testerink: *A.I. for Online Criminal Complaints: From Natural Dialogues to Structured Scenarios*, in prof. of Artificial Intelligente for Justice workshop, collocated at the 22nd European Conference on Artificial Intelligence (ECAI 2016).
5. [Blei et al., 2003] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
6. [Boella et al, 2014] G. Boella, L. Di Caro, A. Ruggeri, L. Robaldo. Learning from syntax generalizations for automatic semantic annotation. Journal of Intelligent Information Systems 43 (2), 231-246
7. [Boella et al, 2016] G. Boella, L. Di Caro, L. Humphreys, L. Robaldo, P. Rossi, L. van der Torre: Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law, Artificial Intelligence and Law, 24 (3).
8. [Deerwester et al., 1990] Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.
9. [Dragoni et al., 2016] Mauro Dragoni, Serena Villata, Williams Rizzi and Guido Governatori, Combining NLP Approaches for Rule Extraction from Legal Documents. In Proceedings of the 1st Workshop on `MIning and REasoning with Legal texts', 2016.
10. [Governatori et al., 2013] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco, “Computing Strong and Weak Permissions in Defeasible Logic,” Journal of Philosophical Logic, vol. 42, no. 6, pp. 799–829, 2013.
11. [Governatori et al., 2009] G. Governatori and S. Sadiq, “The Journey to Business Process Compliance,” in Handbook of Research on Business Process Modeling, J. Cardoso and W. M. P. van der Aalst, Eds. Hershey, New York:, 2009, pp. 426–454.
12. [Governatori, 2005] G. Governatori, “Representing business contracts in RuleML,” International Journal of Cooperative Information Systems, vol. 14, no. 2-3, pp. 181–216, 2005.
13. [Hofmann, 1999] Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.
14. [Huang, 2014] Huang, W.C. “A Patent Document Category System by Using Stanford Parser and

Rough Set Theory“ in Managing access to the internet in public libraries in the UK: the findings of the MAIPLE project, Spacey, Rachel et al. (eds.), ATINER, 2014.

15. [Koehn, 2005] Koehn, Philipp. *Europarl: A Parallel Corpus for Statistical Machine Translation*, MT Summit 2005.
16. [Lesmo, 2007] L. Lesmo. *The Rule-Based Parser of the NLP Group of the University of Torino*. Intelligenza Artificiale, 2(4):46–47, June 2007.
17. [Michael et al, 2014] Michael A., Kirkwood H., Ruqaiya H. *Cohesion in English*. Routledge, 2014.
18. [Robaldo et al, 2012] L. Robaldo, L. Lesmo, D. Radicioni: *Compiling Regular Expressions to Extract Legal Modifications*, In proc. of 25th International Conference on Legal Knowledge and Information Systems (JURIX2012). Amsterdam, 2012.
19. [Schilder, 2005] Schilder, F. *Event extraction and temporal reasoning in legal documents.* Proc. of the 2005 international conference on Annotating, extracting and reasoning about time and events.
20. [Testerink-Bex, 2016] Testerink, B. Floris, B.: *Demo: Natural Language Processing for Online Fraud Scenario Extraction*, demo presented at the 22nd European Conference on Artificial Intelligence (ECAI 2016).
21. [Vico-Calegari, 2015] Vico Horacio, Calegari D. *Software Architecture for Document Anonymization*, Electronic Notes in Theoretical Computer Science (ENTCS), Vol. 314, Issue C, June 2015.
22. [Vila-Yoshino, 1998] Vila, L., Yoshino, H.: Time in automated legal reasoning. Information and Communications Technology Law 7, 173–197 (1998)
23. [Eliantonio, 2013] Mariolina Eliantonio Marta Ballesteros, Rostane Mehdi and Damir Petrovic. Tools for ensuring implementation and application of eu law and evaluation of their effectiveness, July 2013.
24. [Paredes et al, 2008] Luis Polo Paredes, JM Rodrıguez, and Emilio Rubiera Azcona. Promoting government controlled vocabularies for the semantic web: the eurovoc thesaurus and the cpv product classification system. Semantic Interoperability in the European Digital Library, page 111, 2008.
25. [Dumais, 2004] Susan T Dumais. Latent semantic analysis. Annual review of information science and technology, 38(1):188–230, 2004.
26. [Sukhbaatar et al, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In Advances in neural information processing systems, pages 2440–2448, 2015.
27. [Medsker et al, 2001] LR Medsker and LC Jain. Recurrent neural networks. Design and Applications, 2001.
28. [Hochreiter et al, 1997] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735– 1780, 1997.
29. [Weston et al 2014] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. arXiv

preprint arXiv:1410.3916, 2014.

30. [Schrading, 2015] Schrading, J Nicolas, "Analyzing Domestic Abuse using Natural Language Processing on Social Media Data" (2015). esis. Rochester Institute of Technology.

31. [Wyner et al, 2011] Wyner, Adam, and Wim Peters. "On Rule Extraction from Regulations." JURIX. Vol. 11. 2011.

32. [Wyner et al, 2012] Wyner, Adam and Bos, Johan and Basile, Valerio and Quaresma, Paulo. "An empirical approach to the semantic representation of laws." JURIX. 2012.