Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links

Tommaso AGNOLONI [a], Lorenzo BACCI [a], Ginevra PERUGINELLI [a],
Marc van OPIJNEN [b], Jos van den OEVER [b],
Monica PALMIRANI [c], Luca CERVONE [c], Octavian BUJOR [c],
Arantxa ARSUAGA LECUONA [d], Alberto BOADA GARCÍA [d],
Luigi DI CARO [e], Giovanni SIRAGUSA [e]

[a] *Institute of Legal Information Theory and Techniques (ITTIG-CNR)*
[b] *Publications Office of the Netherlands (UBR|KOOP)*
[c] *CIRSFID - University of Bologna*
[d] *General Council of the Judiciary - CENDOJ*
[e] *Computer Science Department - University of Torino*

Abstract. In this paper we present the BO-ECLI Parser, an open framework for the extraction of legal references from case-law issued by judicial authorities of European member States. The problem of automatic legal links extraction from texts is tackled for multiple languages and jurisdictions by providing a common stack which is customizable through pluggable extensions in order to cover the linguistic diversity and specific peculiarities of national legal citation practices. The aim is to increase the availability in the public domain of machine readable references metadata for case-law by sharing common services, a guided methodology and efficient solutions to recurrent problems in legal references extraction, that reduce the effort needed by national data providers to develop their own extraction solution.

Keywords. natural language processing, legal references, case law databases, linked open data

1. Introduction

Among the goals of the European Case Law Identifier (ECLI) established in 2010[1] is the publication of national case-law by courts of European member States via the ECLI Search Engine on the European e-Justice Portal. Besides being uniformly identified, decisions should be equipped with a minimal set of structured metadata describing their main features. Among the (optional) metadata prescribed by the ECLI Metadata Scheme, *references* metadata describe relations of the current document with other legal (legislative or judicial) documents, formally expressed using uniform identifiers (the aforementioned ECLI for case-law, ELI for legislation, national identifiers, CELEX identifiers for European legal documents). These relational metadata are at the same time among the

[1] Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law (CELEX:52011XG0429(01)).

most useful case-law metadata - in that they allow the enhancement of legal information retrieval with relational search - and among the most difficult to have valued, especially for legacy data and for less resourced languages and jurisdictions. While manual reference tagging is an extremely costly procedure - not viable in the public domain and especially to cope with the growing amount of data published in national case law databases - automatic legal reference extraction has been successfully applied in several national contexts [1], [2], [3] despite the complexity of coping with a diversity of styles, variants and exceptions to existing drafting rules and citation guidelines.

Based on an analysis of approaches and existing solutions to the "Linking data" problem [4] and on the results of a survey on citation practices within EU and national Member States' courts [5], the BO-ECLI Parser presented in this work and developed within the EU funded project "Building on ECLI"[2], tackles the problem from a EU-wide multi-lingual / multi-jurisdictional perspective. With a strong commitment to openness (open source software, open data, open formats) the aim is to reduce the effort for national data providers willing to develop their own legal reference extraction solution by sharing a proven methodology and efficient solutions to recurrent problems in legal references extraction.

The BO-ECLI Parser is structured as an architecture of interoperable services (Fig. 1). The core of the extraction process is taken care of by the Parser Engine (Sect. 2). The REST API exposes the results of the reference extraction process as structured interoperable XML and JSON open formats (Sect. 3). Data Services provide access to authoritative repositories of legal references allowing to complement the informations extracted by the Engine (Sect. 4). An extensible User-Interface is also provided for direct user interaction and as a proof of concept of the integration of the different services (Sect. 5).

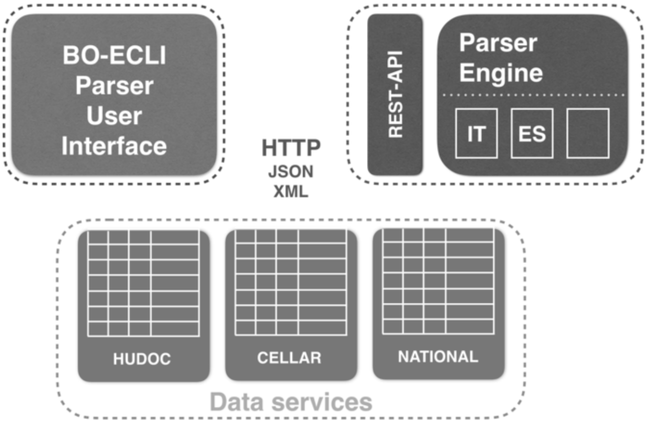

Figure 1. The overall architecture of the BO-ECLI Parser.

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

2. Parser Engine

The BO-ECLI Parser Engine [6] is an extensible framework for the extraction of legal links from case-law texts. It is written in Java and distributed as open source software[3]. It targets citations to both case-law and legislation, expressed as lists of textual features (authority, type of document, document number, date, etc.) or as common names (i.e. aliases). Multiple citations, intended either as citations to more than one partition of a single document or as citations to more than one document issued by a single authority, are also covered and distinct legal references are generated in correspondence to each partition and each document. A distinguishable characteristic of the software consists in the capability to be extended in order to support the extraction process from texts written in different languages or issued within different jurisdictions.
In order to realize such design, two practical steps are required:

- dividing the process of legal link extraction into a generic and customizable sequence of atomic services, following a pipeline pattern;
- defining an annotation system able to convey the work done by each service along the pipeline.

2.1. A pipeline of services

One way to synthesize a generic process of legal link extraction from texts is, first, to divide it into three consecutive phases:

1. the entity identification phase, where the fragments of text that can potentially represent a feature of a citation are identified and normalized;
2. the reference recognition phase, where patterns of identified features are read in order to decide whether they form a legal reference or not;
3. the identifier generation phase, where the recognized legal references are analyzed so that standard identifiers, and possibly URLs, can be assigned to them.

Secondly, within every single phase, a number of different services can be placed, each specialized in absolving one task. For example, within the entity identification phase, there could be a service specialized in the identification of case numbers.

2.2. Annotation system

The BO-ECLI Parser Engine framework defines an internal annotation system to allow every service implementation, especially the ones belonging to entity identification and reference recognition, to save the specific results of their execution directly in the text.
Annotations are used to assign a category (hence, a meaning) to a fragment of text, while, through normalization, annotated fragments of text can acquire a language independent value. For example, the Italian fragment of text *"sent. della Corte Costituzionale"*, meaning a judgment issued by the Italian Constitutional Court, at a certain point along the pipeline, is annotated as follows:

```
[BOECLI:CASELAW_TYPE:JUDGMENT]sent.[/BOECLI] della
[BOECLI:CASELAW_AUTHORITY:IT_COST]Corte Costituzionale[/BOECLI]
```

[3] http://gitlab.com/BO-ECLI/Engine

Thanks to the annotation system, the work of each service is conveyed and shared along the pipeline in a language independent way.

2.3. Service implementation

The implementation of an annotation service belonging to either the entity identification or the reference recognition phase simply consists in a piece of software that analyzes an input text, possibly already enriched with annotations, and produces an equivalent output text, possibly with altered annotations. The default implementations of the annotation services provided by the framework make use of JFlex[4], a well-known lexical scanner generator for Java.

A number of implementations for services that belong to each phase of the legal link extraction process are provided by the framework by default. Typically, a default implementation is supplied when the task that the service is in charge of can be considered language independent, pertains to the European jurisdiction or is common in the European context.

Parties identification: The identification of the names of the parties in a citation should be generally considered as a language dependent task. Nonetheless, the framework provides a default service implementation for the identification of applicants and defendants relying on heuristics based on positioning, upper and lower casing, the *versus* entity and the geographic identification of a country member of the Council of Europe (as a defendant in European Court for Human Rights citations).

Reference recognition: After the entity identification phase, the textual features that can potentially be part of a legal reference are annotated and normalized, hence they can be treated as language independent entities. Although citation practices change from one jurisdiction to another, the framework provides a number of default service implementations for reference recognition that are able to cover the most typical citation patterns and, also, to support multiple citations.

ECLI generation for European Courts: In those cases where a standard identifier can be simply generated as a composition of the features extracted from the textual citation, the framework provides a default service implementation to automatically assign an identifier to a legal reference. This is the case for the generation of ECLI for legal references that have the European Court of Human Rights as the issuing authority, when the type of document, the case number and the date are known.

CELEX generation for European legislation: Another service implementation supplied by the framework for the automatic composition of a standard identifier is used for legislation references to European directives and regulations. For these types of document, when the referred document number and year are known, a CELEX identifier as well as its ELI identifier can be assigned to the legal reference.

3. REST-API and structured reference exchange format

A REST API is wrapped around the Java API of the Engine in order to allow its exposition as a service on the Web via the HTTP protocol and to guarantee interoperability with additional components possibly written in different languages. The Engine REST-API

[4] http://jflex.de

exposes the results of the reference extraction process performed by the Engine as structured XML and JSON open format for their consumption by additional services and for the possible further enrichment and validation of the results of the automatic extraction.

The API response provides, for each text fragment where a citation has been detected, a structured representation of the corresponding reference, listing its attributes along with their normalized values. In case of multiple citations (in the sense described in Sect. 2) a collection of references is returned each associated with the corresponding text fragment.

4. Open Data Services

For those cases where the identifier cannot be computed by the composition of the reference features used in the textual citation, it is mandatory to look-up such standard identifiers (preferably European standard identifiers: ECLI for case-law and ELI and for legislation) by querying reference catalogs. In the BO-ECLI Parser design this is accomplished by reusing existing reference repositories possibly exposed as Open Data on the web accessible via HTTP APIs.

Due to their importance to all national jurisdictions, two data services have been implemented to get standard identifiers of references to case-law issued by the Court of Justice of the European Union (CJEU) and by the European Court of Human Rights (ECHR) for which ECLIs cannot be straightforwardly computed based on the features and numbering typically used in textual citations.

Additionally, national reference repositories can be reused and integrated in order to accomplish national identifiers look-up. Standardizing the access to such metadata repositories through a common layer is among the long term goals of the BO-ECLI Parser framework [7].

5. User-Interface

Though the parser is primarily intended to be used through its API for integration in different systems, an extensible open-source User Interface developed using modern Web technologies (Node.js) is also provided for direct user interaction. The UI interacts with the different Web services through HTTP and provides a proof of concept of their integration. Functionalities are provided to set the input text and parameters and inspect the results extracted by the BO-ECLI Parser in different views: annotated HTML text, tabular view, structured exchange format (JSON) view for developers, *references* metadata according to the official ECLI Metadata Scheme. The UI project is extendable to the needs of the national judiciary for testing or production, e.g. for manual check and validation of the results of the automatic extraction before the deposit in a case law management and publication system. A deployed demo version of the UI is accessible as part of the website of the BO-ECLI project[5].

[5] http://parser.bo-ecli.eu

Conclusions

We presented the BO-ECLI Parser, an open source framework for the automatic extraction of case-law and legislation references from case-law texts issued in the European context. Its architecture is based on the interaction of different interoperable and extensible components. In particular, the Parser Engine provides a framework where national extensions can be developed and plugged in order to add support for the extraction process from texts written in different languages or issued within different jurisdictions. By defining and providing a complete stack for legal links extraction, the implementation of a national extension is guided and straightforward and the effort needed for the development of a fully functional national extractor is considerably reduced. Along with the framework project, a Template project has been developed in order to facilitate and encourage the adoption of the software for the extraction of legal links in new languages and jurisdictions. Two concrete national extensions have been developed so far by different teams to support the extraction from Italian and Spanish case-law texts, proving both the feasibility and the straightforwardness of the whole approach. The BO-ECLI Parser software projects, their code and documentation are hosted on the GitLab software development platform[6].

Acknowledgement

This publication has been produced with the financial support of the Justice Programme of the European Union. The contents of this publication are the sole responsibility of the authors and can in no way be taken to reflect the views of the European Commission.

References

[1] Lorenzo Bacci, Enrico Francesconi and Maria Teresa Sagri. A Proposal for Introducing the ECLI Standard in the Italian Judicial Documentary System. In *Proceedings of the 2013 Conference on Legal Knowledge and Information Systems: JURIX 2013: The Twenty-sixth Annual Conference* pp. 49-58 IOS Press Amsterdam (NL), 2013.

[2] Marc van Opijnen, Nico Verwer and Jan Meijer. Beyond the Experiment: The Extendable Legal Link Extractor. In *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, June 8-12 2015 held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL) San Diego, CA, USA. Available at SSRN: https://ssrn.com/abstract=2626521.

[3] A. Mowbray, P. Chung and G. Greenleaf. A free access, automated law citator with international scope: the LawCite project. *European Journal of Law and Technology* **7** (3), 2016. Available at: http://ejlt.org/article/view/496/691 .

[4] Tommaso Agnoloni and Lorenzo Bacci. *BO-ECLI project deliverable D2.1 Linking Data - analysis and existing solutions*, 2016. Available at: http://bo-ecli.eu/uploads/deliverables/DeliverableWS2-D1.pdf .

[5] Marc van Opijnen, Ginevra Peruginelli, Eleni Kefali and Monica Palmirani. *On-line Publication of Court Decisions in the EU - Report of the Policy Group of the Project 'Building on the European Case Law Identifier'*, 2017. Available at: http://bo-ecli.eu/uploads/deliverables/Deliverable%20WS0-D1.pdf .

[6] Tommaso Agnoloni, Lorenzo Bacci and Marc van Opijnen. BO-ECLI Parser Engine: the Extensible European Solution for the Automatic Extraction of Legal Links. In *Proceedings of the 2nd Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, June 16 2017 held in conjunction with the 2017 International Conference on Artificial Intelligence and Law (ICAIL) London, UK.

[7] Marc van Opijnen, Monica Palmirani, Fabio Vitali, Jos van den Oever and Tommaso Agnoloni. Towards ECLI 2.0. In *CeDEM17 Proceedings of the International Conference for E-Democracy and Open Government*, May 17-19 2017. Danube University Krems, Austria.

[6] https://gitlab.com/BO-ECLI