Building a corpus for the anonymization of Romanian jurisprudence

Vasile Păiș and **Dan Tufiș** and **Elena Irimia**
and **Verginica Barbu Mititelu**
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
Bucharest, Romania
vasile@racai.ro

Abstract

Access to jurisprudence is of paramount importance both for law professionals (judges, lawyers, law students) and for the larger public. In Romania, the Superior Council of Magistracy holds a large database of jurisprudence from different courts in the country, which is updated daily. However, granting public access to it requires its anonymization. This paper presents the efforts behind building a corpus for the anonymization process. We present the annotation scheme, the manual annotation methods, and the platform used.

1 Introduction

The astonishing advancement of Machine Learning (ML) and Artificial Intelligence (AI) during the last decade has generated a global rush for collecting more and diverse data, as clean as possible, with an eye to the General Data Protection Regulation (GDPR)[1] compliance. Large language models (LLMs), fueling the most successful AI applications, are built from data collected from various sources, the web being the most frequent one, but not the only one. When access to the data is open to the public, according to global GDPR requirements, any personal/private information must be hidden.

The procedure of hiding/obscuring/obfuscating personal data in documents released to the public is known as documents anonymization. It has to be performed so that the remaining context could not unveil the purposely hidden information. This is, generally speaking, a hard task, but, for specialized texts/language, it gains in accuracy.

The judicial systems all over the world are under the scrutiny of people, who naturally claim the right to have access to information on the decisions affecting their lives. Transparency of judicial decisions, as well as consistency of national courts decisions among themselves and with the international practices and recommendations are highly sensitive topics.

To tackle these issues, the Council of Europe is implementing the project "Fostering transparency of judicial decisions and enhancing the national implementation of the European Court on Human Rights" (TJENI)[2], which aims to improve the transparency and consistency of national judicial decisions, to strengthen the quality of their judicial decision-making and to streamline information on human rights jurisprudence to national judiciaries. The protection of human rights and the rule of law are strengthened by transparency of the judicial process, increasing the consistency of national courts decisions with European human rights and rule of law standards. These objectives can be supported through the publication of court decisions, which requires their prior anonymization. This can be done by applying specific technical solutions meant to automate the preparation of the documents for publication.

Romania is the only TJENI beneficiary that publishes decisions of all courts in Romania through the portal specifically developed by the Superior Council of Magistracy (SCM), with the exception of the High Court of Casation and Justice, which maintains its own case law database. The case law database of the SCM, the official beneficiary of our project, contains all court decisions from criminal, civil, commercial and administrative case types. The only exceptions from publication are documents marked in ECRIS (the case management system) as confidential, such as the judgements or other decisions related to minors, sexual harassment, divorces/family matters, decisions acknowledging mediation settlements, verdicts on offences as treason, espionage, rape and child pornography

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[2] https://www.coe.int/en/web/national-implementation/tjeni

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 71–76

and upon a request of a party in the proceeding. So far, more than 40 million decisions have been published since 2011 in the Romanian case law database and the database is updated daily. Initially, decisions were anonymized before publication by means of regular expressions. A new anonymizer, developed based on this work, more accurate and much faster, will replace the previous one.

The new tool is being built with out-of-the-box scalability, by means of parallelism and containerization mechanisms, allowing for high-performance processing of an increasing number of documents. Furthermore, the new system employs state-of-the-art LLMs, such as Romanian language BERT (Devlin et al., 2019) models, for detecting named entities (NEs) that need to be anonymized, thus increasing the recognition performance. The corpus described in this work will be used for training the tool. The system still employs dictionaries and regular expression lists for certain types of NEs, that are particularly suited for such recognition processes (like personal identification codes, vehicle registration, email addresses).

The rest of this paper is structured as follows: Section 2 presents related work, Section 3 describes the annotation scheme, Section 4 presents the manual annotation process in the RELATE platform, Section 5 introduces preliminary statistics on the corpus, and we conclude in Section 6.

2 Related work

The anonymization process implies named entity recognition (NER). However, not all NEs require anonymization (as detailed in Section 3). Given this link between anonymization and NER, this section will cover corpora built either for NER or for anonymization in the legal domain. Plamondon et al. (2004) admit that anonymization of court decisions presupposes proper identification of more than just person names, while too much anonymization reduces the text readability and usability.

Trias et al. (2021) are concerned with the identification of lawyer names in historical legal text. They acknowledge problems arising from using nicknames or initials instead of complete names. Leitner et al. (2019) perform fine-grained NER on a corpus (Leitner et al., 2020) of German legal documents. The corpus was constructed from already anonymized court decisions, thus affecting the NEs belonging to personal data. Their specification contains 19 fine-grained NE classes. Au et al. (2022) constructed the E-NER dataset, based on legal company filings available from the US Securities and Exchange Commission's EDGAR data set, containing 7 NE classes.

Legal-ES (Samy et al., 2020) is a large Spanish corpus covering different types of legislative, administrative and jurisprudential texts. Kalamkar et al. (2022) describe a NER corpus for Indian court judgements, covering 14 NE classes.

Considering the Romanian language, Păiș et al. (2021a) constructed the LegalNERo corpus, covering persons, locations, organizations, time expressions, and legal references (5 NE classes). The corpus consists of a manually annotated subset of the larger MARCELL-Ro corpus (Tufiș et al., 2020; Váradi et al., 2020), containing legal domain texts, primarily legislation. This corpus, without the legal reference annotations (which do not have a corresponding class for the current project), could be used to augment the jurisprudence corpus described in the rest of this paper.

All these papers admit the necessity to annotate more types of entities in corpora from the legal domain. As detailed in Section 3, we also considered the annotation of more types of entities, as well as the necessity of their annotation (see Table 1 and the discussion about the data it shows).

3 Annotation scheme

Our annotation scheme is based on NE labels commonly used in the NER field (like LOC, ORG, PER and DATE) and is extended to accommodate entities specific to the legal domain, especially the jurisprudence context of our project (with labels like ECLI, CASE, DECISION, etc.). The scheme contains 17 entity types (see Table 2 for the complete list of labels), in accordance with the anonymization requirements defined in the Decision of the Section for Judges of the Superior Council of the Magistracy no. 998/17.03.2022 [3].

The guidelines for using the annotation scheme were adapted to the anonymization task, which is different from NER in the sense that not all NEs need being anonymized.

The anonymization task particularities resulted in a classification of the entity types according to their consistency of annotation: all occurrences of certain entity types (e.g., ECLI, EMAIL, CUI, IBAN, etc.) in the target documents are annotated,

[3] http://old.csm1909.ro/csm/linkuri/02_05_2022__105390_ro.pdf

while for other types (such as PER, DATE, LOC, ORG, etc.), the decision to annotate is based on the type of reference those entities have (see below). To have a sense of the degree of distinction between the anonymization and the NER task in our framework, we randomly selected 20 documents from the set already manually annotated for anonymization and supplemented the annotation to include all the NE occurrences corresponding to our entity types of interest. Table 1 shows the number of annotated entities in the anonymization task and the number of those annotated in the NER task. It is easily visible that the total number of entities is almost double (794 vs. 1,479) in the NER task as compared to the anonymization one.

	Anonymization Task	**NER Task**
PER	312	318
DATE	26	359
LOC	70	147
ORG	113	376
TOTAL	794	1,479

Table 1: The number of annotated entities for NER task vs. Anonymization task in the same documents. Entity types whose number of annotated occurrences is identical in the two tasks are not detailed in the table, but the total includes them.

PER entities, which include human names, surnames and nicknames, are to be annotated in the vast majority of situations, regardless of the person's role in the trial (petitioner, respondent, convict, witness, judge, clerk, etc.); yet, there is one exception to this rule: when the person's name is cited in connection with a European Court of Human Rights case, which is, by its nature, public, and does not need to be anonymized (examples from our corpus: "cauza Salabiaku c. Franței" ("the case of Salabiaku v. France"), "cauza Västberga Taxi Aktiebolag și Vulic c. Suediei" ("the case of Västberga Taxi Aktiebolag and Vulic v. Sweden"). Person names also occur preceded by some phrases indicating a legal entity: e.g. "birou de avocatură" ("law office"), "cabinet medical" ("medical office"); in this case they are annotated as ORG.

DATE entities (used to annotate time expressions) are in the opposite situation, i.e. they must be annotated on a few specific occasions, namely when they refer to the date of birth of a person; the jurisprudence documents are rich in occurrences of DATE NEs which are not to be anonymized: dates when the trial takes place, when a decision is made, a document is issued, etc.

LOC entities are anonymized only if otherwise they could disclose the identity of persons or organizations. Therefore, they are always anonymised when referring to residence, place of birth, headquarters and buildings/land in possession. Locations where accidents and events take place are not normally annotated, but there are exceptions to this rule: e.g., when the event takes place in a small town/village or a specific geographical location that is very close to the individual residence and whereabouts information could endanger the anonymity of the parties. The decision to annotate such occurrences is made for each case separately.

ORG entities include all groups defined by a formal organisational structure, whether public or private. While private organizations are always anonymized, public organisations are annotated only when they are parts in the trial. By their nature, the jurisprudence documents abound in ORG named entities, with many of them representing law institutions that do not require anonymization. This is reflected in the tripling of ORG entities in our evaluation trial (see Table 1).

All remaining NE types (see Table 2 for a complete list of labels) are always annotated. INITIALS only refer to occurrences of initials instead of signatures (of the judge and clerk) at the end of the documents. Other types of abbreviations occurring in documents are annotated as the NE type they abbreviate: e.g., companies initials are annotated as ORG, person initials are annotated as PER. Court decisions, CNP, ID, EMAIL, ECLI, CUI, IBAN, NCAD and AUTO are annotated without any exception. Some of them have a homogeneous format (for example, EMAIL is easily recognized by the presence of @ and of a dot), while others can take different forms. AUTO is such an example: it is used to annotate both Romanian plate numbers and foreign ones, which can have a different format; even in Romania, plate numbers belonging to official institutions cars or provisional plate numbers have different formats from the common ones. Moreover, there are cases when the same number is typed differently throughout the same decision (e.g., "GH13ABC" – "GH 13 ABC" – "GH-13-ABC"). The AUTO label is also used for other vehicle identification numbers, like chassis series, which have a different format than the plate numbers. Thus, the annotation of entity types that does not depend on the semantic context of occurrence can also be problematic at times, when formats are

heterogeneous inside the same type.

4 Manual annotation in the RELATE platform

RELATE (Păiș et al., 2020) is a modern platform incorporating a large number of tools (Păiș, 2020) for processing the Romanian language. It was previously used for automatic annotation of large corpora, such as the MARCELL (Váradi et al., 2020; Tufiș et al., 2020) legislative corpus and the CURLICAT (Váradi et al., 2022) corpus and for creating Romanian language named entity corpora, such as MicroBloggingNERo (Păiș et al., 2022) and LegalNERo (Păiș et al., 2021b).

For the purpose of this work, we use a number of RELATE platform's modules, including: corpus management, manual annotation and basic language annotation resource kits. For security reasons, with regard to data access, the modules were deployed in the secure network of the Superior Council of Magistracy, and the web interface was made available to annotators via VPN connections. For manual annotation, the RELATE platform integrates the BRAT Rapid annotation tool (Stenetorp et al., 2012), connected to the platform's corpus management component. Annotators are shown one document at a time and must select, using the mouse, each NE text span. The platform remembers the last document worked on by each annotator, as well as documents not finalized, allowing smooth transition between documents. At this stage, the NEs are only marked in the documents and no anonymization takes place, as the corpus is intended for training automatic processes later. Following the span-level annotation, documents are processed using UDPipe (Straka et al., 2016) with a custom model (Păiș et al., 2021) trained on the Romanian RoRefTrees (RRT) corpus version 2.7 (Barbu Mititelu et al., 2016), available in the Universal Dependencies project. The resulting tokenized version is automatically aligned with the span-level NE annotations, using a BIO (begin/inside/outside) annotation format.

5 Preliminary corpus statistics

The manual annotation task (which is currently in progress and done by 38 annotators) is intended to cover 1,500 documents and double annotation was taken in consideration for inter-annotator agreement (IAA) analysis. At the moment, based on 5,563,617 tokens from documents doubly annotated, the IAA score, computed using Cohen's Kappa, is 0.94. According to Landis and Koch (1977), a Kappa value greater than 0.81 is indicative of an "almost perfect" agreement. The annotators are primarily judges with experience in anonymization requirements for legal documents, working under the coordination of the Superior Council of Magistracy. This accounts for the high agreement score. Two additional annotators with experience in creating annotated corpora were involved in order to better understand the data and clarify disagreements. Throughout the annotation process, periodic discussions took place to clarify any problems.

Entity Class	# entities
AUTO (car plates)	218
CASE (trial case number)	2,028
CNP (personal numeric code)	291
CUI (commercial unique identifier)	82
DATE	1,444
DECISION (trial decision number)	2,532
DOC(ument)**NUMBER**	3,151
E(uropean)**C**(ase)**L**(aw)**I**(dentifier)	157
EMAIL (address)	3
IBAN	2
I(dentification)**D**(ocument)	73
INITIALS	1,149
LOC(ation)	2,370
NCAD (land registry number)	167
ORG(anization)	3,662
PER(son)	19,557
PHONE (number)	28
TOTAL	36,914

Table 2: Preliminary statistics on 594 unique annotated files with 5,563,617 tokens

As shown in Table 2, the documents are very rich in personal names and the vast majority of them have to be annotated as entities to be anonymised (19,557 PER entities); on the other end, entities like EMAIL (3 occurences) and IBAN (2 occurences) are very rare.

6 Conclusion and future work

This paper introduced the work carried out for creating a corpus for the purpose of anonymizing the Romanian jurisprudence. It is a very challenging task, with its own peculiarities when it comes to automatic processing. Even though it is suitable for an NER approach, the fact that only entities

requiring anonymization (not all the entities) are annotated makes it difficult to use readily available NER applications. Thus, a combination of different methods are explored, including a combination of algorithms based on traditional techniques (dictionary, regular expressions) and large language models. Even though completely annotating the corpus for NEs and marking those that require anonymization would have enabled additional uses for the corpus, given the annotators experience in anonymization (and not NER or other corpus building activities) it was decided to focus only on the anonymization task.

System development, including algorithm's implementation, is realized open source[4], but the corpus itself cannot be publicly released, due to the sensitivity of the information. We are also considering releasing pre-trained models (when this would not compromise privacy) and anonymized corpus samples. The end-result of the project, the anonymized jurisprudence, will be available from the ReJust portal[5].

Acknowledgements

This work was performed within the project "System for the anonymization of Romanian jurisprudence (SAROJ). The SAROJ project was carried out with funding from the Council of Europe, within the project "Foster Transparency of Judicial Decisions and Enhancing the National Implementation of the ECHR (TJENI)" funded by the EEA and Norway Grants Fund for Regional Cooperation. The views expressed herein can in no way be taken to reflect the official opinion of the Council of Europe.

References

Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Verginica Barbu Mititelu, Radu Ion, Radu Simionescu, Elena Irimia, and Cenel-Augusto Perez. 2016. The romanian treebank annotated according to universal dependencies. In *Proceedings of The Tenth International Conference on Natural Language Processing (HrTAL2016)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham. Springer International Publishing.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Vasile Păiș, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, and Dan Tufiș. 2021. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401.

Vasile Păiș, Radu Ion, and Dan Tufiș. 2020. A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France. European Language Resources Association.

Luc Plamondon, Guy Lapalme, and Frédéric Pelletier. 2004. Anonymisation de décisions de justice. In *Actes de la 11ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 159–168, Fès, Maroc. ATALA.

Vasile Păiș. 2020. Multiple annotation pipelines inside the relate platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75.

Vasile Păiș, Maria Mitrofan, Verginica Barbu-Mititelu, Elena Irimia, Carol Luca Gasan, Roxana Micu, Laura Marin, Maria Dicusar, Bianca Florea, and Ana Badila. 2022. Romanian micro-blogging named entity recognition (MicroBloggingNERo).

[4] https://github.com/racai-ai/saroj/
[5] https://rejust.ro

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021a. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. 2021b. Romanian Named Entity Recognition in the Legal domain (LegalNERo).

Doaa Samy, Jerónimo Arenas-García, and David Pérez-Fernández. 2020. Legal-ES: A set of large scale resources for Spanish legal text processing. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 32–36, Marseille, France. European Language Resources Association.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. 2021. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 172–179, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. 2020. Collection and annotation of the romanian legal corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2766–2770, Marseille, France. European Language Resources Association.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. The marcell legislative corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3754–3761, Marseille, France. European Language Resources Association.

Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartlomiej Nitoń, Piotr Pęzik, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, and Andraž Repar. 2022. Introducing the curlicat corpora: Seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Language Resources and Evaluation Conference*, pages 100–108, Marseille, France. European Language Resources Association.