LegalNERo: A linked corpus for named entity recognition in the Romanian legal domain

Vasile Păias, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onut

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Romania
E-mail: vasile@racai.ro

Abstract. LegalNERo is a manually annotated corpus for named entity recognition in the Romanian legal domain. It provides gold annotations for organizations, locations, persons, time expressions and legal resources mentioned in legal documents. Furthermore, GeoNames identifiers are provided for location entities, when linking was possible. The resource is available in multiple formats, including span-based, token-based and RDF. The Linked Open Data version, in RDF-Turtle format, is available for both download and interrogation using a SPARQL endpoint.

Keywords: Named entity recognition, Linguistic linked data, Romanian language, Corpus

1. Introduction

Named entity recognition is the task of identifying named entities in text [45], such as persons, locations, organizations, time expressions, proteins, etc. Starting in 1995, within the MUC-6 conference [15], there have been periodic tasks on various aspects of named entity recognition, focusing on different entity types. For example, for the CoNLL-2003 shared task on language-independent named entity recognition [41], named entities were considered "phrases that contain the names of persons, organizations and locations". However, this limited approach is not suitable for every domain. In this context, the biomedical domain, a number of works have addressed entities such as genes, proteins, diseases [18], cell types [38], and chemicals [14], [20].

In the legal domain, the TREC conference had a dedicated track [8] administered by NIST [1] for evaluating the application of Information Retrieval (IR) methods to e-discovery in the context of the U.S. civil litigation from 2006 until 2011 [27]. The Competition on Legal Information Extraction and Entailment (COL-IEE) [21] run over multiple editions allowed further exploration of tools and algorithms for information extraction in the legal domain.

In the context of the EU project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL) [2], a large comparable corpus of legal documents for seven languages was created [44]. Comparable corpora can be exploited for improving machine translation [26] and parallel sentences can be identified automatically, potentially containing named entities. The Romanian sub-corpus [42], as well as the other MARCELL corpora, was split at the sentence and token levels, lemmatized, and annotated at the token level. Annotations comprise part-of-speech tags, dependency parsing, named entities and finally the corpus was enriched with IATE terms and EUROVOC descriptors. All the annotations in the large Romanian MARCELL corpus were realized using automatic processes.

Existing Romanian named entity corpora include: RONEC [11], Romanian TimeBank [12] and SiMoNERo [1]. The RONEC corpus contains 26,377 named entities, belonging to 16 different classes. The Romanian TimeBank corpus is an annotated parallel corpus for temporal information. This corpus contains 26,635 temporal named entities such as events,
instances, and signals. SiMoNERo is a gold standard corpus for the biomedical domain, manually annotated with four types of domain-specific named entities. SiMoNERo has 14,133 named entities distributed in 4,987 sentences. In this corpus, the NEs are in BIO format. All these corpora contain entities such as organizations, persons, locations and also time expressions and biomedical entities. Nevertheless, none of these corpora contains legal texts or legal entities.

This paper presents a manually annotated corpus comprising a subset of documents from the MARCELL Romanian corpus, with named entities in the legal domain. We considered the classical entity types (organizations, persons, locations) and time expressions as they appear in legal documents and added a new entity type in the form of legal references to documents (such as laws, government decisions, orders, etc.).

The paper is structured as follows: in Section 2 we present related work; in Section 3 we introduce the annotation process of the corpus; Section 4 describes different aspects of the corpus, such as the annotation levels, the representation of the linked data and statistics of the corpus; Section 5 considers aspects regarding the usage of the RDF version of the corpus; Section 6 presents real use cases for the corpus; and finally, we conclude in Section 7.

2. Related Work

One of the first papers to discuss named entity recognition in the legal domain is that of Dozier et al. [10]. The authors explore named entity recognition and resolution in legal documents such as US case law, depositions, pleadings and other trial documents. The types of entities include judges, attorneys, companies, jurisdictions, and courts.

Cardellino et al. [4] explored using the LKIF\textsuperscript{3} ontology [17] further mapped to the YAGO\textsuperscript{4} ontology [40] to train a named entity recognizer, classifier and linker. The resulting system is then applied to a corpus comprising judgements of the European Court of Human Rights. The authors recognize that in the legal domain named entities are also names of laws, typified procedures and even concepts. Furthermore, when dealing with human annotators they observe that the classes and subclasses of Document, Organization and Person were the most consistent across annotators.

Glaser et al. [13] explored the suitability of named entity recognition systems in the case of legal contracts. The proposed entity classes are person, organization, location, date, money value, reference, and other. The "reference" entity is based on the work of [23], where references to legal norms are considered.

Leitner et al. [24] introduced a German legal named entity corpus comprising seven coarse-grained classes that can be expanded into 19 fine-grained classes. In this case, a "person" entity can be classified into a regular person, a judge or a lawyer. Similarly, a "legal norm" entity can be further expanded into law, ordinance or European legal norm.

In the MARCELL project, Romanian named entities were identified using a general-purpose tool [30], available at that time for the Romanian language, that was not adapted to the legal domain, allowing only entities such as organization, persons, locations and also time expressions. The tool was not trained on any legal texts, but since no legal-domain tool was available it was used on this corpus. Thus, even at that time the need for a legal-domain NER corpus and system became apparent, which later led us to the development of the LegalNERo corpus.

3. Annotation process

Annotation was performed by five human annotators, under the supervision of two senior researchers at the Institute for Artificial Intelligence "Mihaile Drăgănescu" of the Romanian Academy (RACAI)\textsuperscript{5}. Annotators followed specific guidelines\textsuperscript{6}, inspired in part by the Linguistic Data Consortium (LDC) guidelines for annotation of named entities\textsuperscript{7}.

We considered five classes: person (PER), location (LOC), organization (ORG), time expressions (TIME) and legal document references (LEGAL). For person entities, we considered only people names. Titles and honorifics present in text near a person name were not included in the entity. Organizations must have some formally established association. Typical examples are businesses, government units and political parties. Lo-

\textsuperscript{3}https://github.com/RinkeHoekstra/lkif-core
\textsuperscript{4}https://yago-knowledge.org/
\textsuperscript{5}https://www.racai.ro/en/
\textsuperscript{6}https://relate.racai.ro/resources/legalnero/legalnero_annotation_guide.pdf
\textsuperscript{7}https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf
sections are defined on a geographical basis and include countries, cities and other geographical areas. References are introduced similar to [23] and the coarse-grained class of [24], without additional sub-classes. Thus, they are references to legal documents such as laws, ordinances, government decisions, etc. Even though we only annotated the legal reference coarse-grained class, most of these entities can be mapped to fine-grained classes using automated processes, employing other linked data resources, as exemplified in Section 5.

Each annotator was given instructions on how to annotate the documents and then annotated a single document (outside of the corpus). We then discussed any issues or questions the annotators had. Subsequently, a collection of 100 documents was attributed to each annotator. Thirty documents (out of the 100) were also shared with two other annotators, allowing us to later compute inter-annotator agreement (IAA). Since the 30 shared documents were mostly situated in the first part of the corpus, this aspect was hidden from the annotators, thus making them focus their attention on all the files. Throughout the annotation process, we held periodic meetings to discuss any issues.

Corpus and account management for the annotators was realized through the RELATE platform [28]. Actual annotation was handled using the BRAT \(^8\) annotation tool [39], integrated into the RELATE platform. This allowed the annotators to view one document at a time, select the identified entity with the mouse and then associate an entity type with the selected text span.

After the annotation process ended, we were able to compute inter-annotator agreement between each pair of annotators, using Cohen’s Kappa measure. This was accomplished at the token level and led to an average Kappa of 0.87. Following this result, we further investigated the differences and we were able to detect some recurring mistakes with some of the annotators, such as the inclusion of indicative words in the entities (for example “orasul București” / “the city of Bucharest” instead of just “București” / “Bucharest”). An automatic script was created to correct these types of mistakes.

Finally, we constructed an application to manually merge the common annotations into a single file. For each entity, the application shows all the other entities overlapping the same span (if they exist) and allows the user to select the entities that go in the final merged file. The application further makes this process easy by highlighting entities found by multiple annotators.

Once all the common annotations were merged we recomputed Cohen’s Kappa measure between the merged corpus and each annotator. This produced an average Kappa of 0.89 and we consider this to be the final result. According to Landis and Koch [22] a value of Kappa greater than 0.81 is indicative of an “almost perfect” agreement between the annotators. The remaining disagreements account for mistakes made by individual annotators, such as missing an entity or a sub-entity. This is particularly reflected in potentially ambiguous situations such as the legal reference “Regulamentul CE nr. 765/2008” (en. “Council Regulation EC No 765/2008”). In this case certain annotators identified “CE” as an organization sub-entity, while others did not. In a few cases the end-of-sentence punctuation coincides with the dot indicating an abbreviation. Some annotators included the punctuation in the organization entity abbreviation, while others considered it to be sentence punctuation and did not include it (for example: ”S.R.L.” vs “S.R.L” or “AFER.” vs “AFER”).

4. Corpus description

4.1. Annotation levels

Raw text files were extracted from the Romanian part of the MARCELL corpus. They contain national legislation gathered by crawling from the public Romanian legislative portal. As described in [44], the texts were extracted from the original HTML format and converted into TXT files. For the purposes of constructing the LegalNERo corpus, we selected 370 documents of similar size, issued in the last two years (2020–2021). We also performed an initial check to ensure that the files contain correct Romanian characters (with diacritics) and do not contain tables or other structures that may impact the annotations.

As described in Section 3, annotation was performed using the BRAT tool integrated into the RELATE platform. Thus, the primary annotation output is represented by BRAT-specific files. Each line contains an entity ID, followed by the entity type, the text span (start and end characters) and the actual text. This annotation format allows for multiple annotations in overlapping spans.

\(^8\)https://brat.nlplab.org/index.html

\(^9\)http://legislatie.just.ro/
We used UDPipe\textsuperscript{10} on the text files for automatic operations such as tokenization, lemmatization, part of speech tagging and dependency parsing. The resulting files were in CoNLL-U format\textsuperscript{11}. This format can be extended with additional columns by following the CoNLL-U Plus guidelines and adding in a special metadata line the description of the new columns. Using this approach we added a new column "RELATE:NE" (the 11th column) for named entity annotations. We mapped the identified annotation text spans to tokens using a BIO notation format [37]. This implies that each token has an additional annotation with the associated entity, prefixed with one of "B-" (for entity beginning) or "I-" (for token inside the entity). Tokens that are not part of any entity are annotated with "O" ("outside").

The use of the BIO annotation scheme means that there is no direct support for overlapping entities. A token is associated with a single entity type. Therefore, we created two separate token-based annotations, stored in two corresponding folders: one for storing all the entity types, without embedded entities, considering only the largest text spans, and another for storing only person, organization, and location entities and time expressions.

Provision of the two CoNLL-U Plus folders means the corpus can be easily used either for legal domain annotations (considering the legal references) or for general annotations (considering the other entity types).

Initial annotations (BRAT and CoNLL-U Plus) were converted to RDF format, specific to applications exploiting linked data. This increases the usability of the corpus and allows analysis of the corpus using RDF queries and linking to external databases.

Location entities can be resolved to places in the real world using geographical databases, such as GeoNames\textsuperscript{12}. We considered this to be a useful property, therefore we added an additional annotation level considering the GeoNames identifiers. The annotation is available in both CoNLL-U Plus files (column 12, "RELATE:GEONAMES") and in the RDF representation. Figure 1 gives an example of the span-based annotation in ann format. Figure 2 gives an example of the token-based annotation in CoNLL-U Plus format.

4.2. Linked data representation

Already having the text span annotations (in BRAT format) and the token-based annotations (in CoNLL-U Plus format) we were faced with the problem of designing a schema useful for linked data applications. First, we considered the CoNLL-RDF representation [6],[7]. It directly translates from tab-separated CoNLL format to RDF by employing the prefix "conll" together with the column name. It further associates a token representation with the NLP Interchange Format (NIF) ontology [16], by declaring it as a "nif:Word" element, linked to a "nif:Sentence".

We further investigated the POWLA [5] ontology. This was also used by [7] complementary to the NIF ontology. Unlike other approaches, POWLA is not tied to a specific selection of annotation layers, but it is designed to support any kind of text-oriented annotation. For this purpose, POWLA allows specifying "document layers" that contain the actual annotations. This is very similar to our situation, where we have an annotation layer comprising the text spans associated with entities (corresponding to the BRAT format) and the token-based annotations (corresponding to the CoNLL-U Plus format).

For the named entity annotations, we employed the NERD ontology [35]. It was previously mentioned [36] that NERD can be used together with the NIF ontology. It provides classes such as "nerd:Location", "nerd:Person", "nerd:Organization" and "nerd:Time" that can be used for the corresponding entities and time expressions. Nevertheless, there is no direct specification for legal references.

The European Legislation Identifier (ELI) ontology provides a descriptive framework for structuring metadata of legislative resources and publishing them as linked data. Its primary purpose is to describe relationships between national and European legislative resources. It provides the "eli:LegalResource" class

\textsuperscript{10}https://ufal.mff.cuni.cz/udpipe
\textsuperscript{11}https://universaldependencies.org/format.html
\textsuperscript{12}https://www.geonames.org/
which is defined as a work in a legislative corpus, applying to acts that have been legally enacted (whether or not they are still in force).

The GeoNames database integrates geographical data such as the names of places in various languages, elevation, population and others from various sources. According to the information available on the website\(^1\), it contains over 25 million geographical names and consists of over 11 million unique features, including 4.8 million populated places and 13 million alternate names. We linked location entities with the GeoNames database by using the feature identifiers associated with each GeoNames feature. The annotation was performed automatically and then manually validated. We used the "gn:Feature" attribute in the corresponding entity or token annotation.

Table 1 presents the vocabularies used in the corpus. The key concepts and relationships expressed in the dataset are visualized in Figure 3. Some of the vocabularies from Table 1 were used only as part of metadata specification, therefore they do not appear in the diagram. We used the graphical ontology editor OWLGrEd [3] to construct the diagram shown in Figure 3. A complete example is included in Appendix A.

The corpus comprises multiple documents, represented as "powla:Document" elements. Each document is organized into three layers ("powla:DocumentLayer"), corresponding to sentences, tokens and named entity text spans. Each object’s layer is indicated by the "powla:hasLayer" attribute and the layer is linked to a document using the "powla:hasDocument" attribute. Tokens are linked to the corresponding sentences, using the "nif:sentence" and "powla:hasParent" attributes. Furthermore, the order of tokens in a sentence can be maintained using the "powla:next" and "powla:previous" attributes. All the CoNLL-U Plus information is available at token level, including word form ("conllu:FORM"), lemma ("conllu:LEMMA"), universal part-of-speech ("conllu:UPOS"), language-specific part-of-speech ("conllu:XPOS"), morphological features ("conllu:FEATS"), dependency information ("conllu:HEAD" and "conllu:DEPREL"), named entity type ("conllu:NE"), and GeoNames identifier ("gn:Feature"). In order to reduce the file size of the dataset, empty attributes are not present in the representation. For example, the first word in a sentence will not have a "powla:previous" attribute.

Since the corpus is available in multiple representations (raw text, span-based annotations, token-based annotations and linked data RDF), we follow each facet and present the corresponding statistics. Table 2 presents the general corpus statistics. There are 265,335 tokens distributed in 370 documents. The average length of a sentence is 32.02 tokens, which is above the 16.06 tokens/sentence, the average sentence length in ROMBAC [19], a balanced Romanian corpus, containing legal, news, fiction, biographical and medical texts.

Table 3 presents the distribution of the annotated tokens conllup files of the corpus. It can be seen that legal documents references class (LEGAL) contains 2,851 organizations (ORG) and 3,301 time (TIME) expressions. This format of the corpus also contains 1,411 GeoNames identifiers linked with the locations (LOC),

\(^1\)https://www.geonames.org/about.html
Table 1
Used vocabularies

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Name</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>nif</td>
<td>NLP Interchange Format (NIF)</td>
<td><a href="http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#">http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#</a></td>
</tr>
<tr>
<td>powla</td>
<td>POWLA Ontology</td>
<td><a href="http://purl.org/powla/powla.owl#">http://purl.org/powla/powla.owl#</a></td>
</tr>
<tr>
<td>nerd</td>
<td>NERD Ontology</td>
<td><a href="http://nerd.eurecom.fr/ontology#">http://nerd.eurecom.fr/ontology#</a></td>
</tr>
<tr>
<td>conllu</td>
<td>CoNLL-U tabular format</td>
<td><a href="https://universaldependencies.org/format.html#">https://universaldependencies.org/format.html#</a></td>
</tr>
<tr>
<td>conllup</td>
<td>CoNLL-U Plus format</td>
<td><a href="https://universaldependencies.org/ext-format.html#">https://universaldependencies.org/ext-format.html#</a></td>
</tr>
<tr>
<td>eli</td>
<td>European Legislation Identifier (ELI)</td>
<td><a href="http://data.europa.eu/eli/ontology#">http://data.europa.eu/eli/ontology#</a></td>
</tr>
<tr>
<td>gn</td>
<td>GeoNames</td>
<td><a href="http://www.geonames.org/ontology#">http://www.geonames.org/ontology#</a></td>
</tr>
<tr>
<td>rdf</td>
<td>RDF</td>
<td><a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a></td>
</tr>
<tr>
<td>rdfs</td>
<td>RDF Schema</td>
<td><a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a></td>
</tr>
<tr>
<td>owl</td>
<td>OWL</td>
<td><a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a></td>
</tr>
<tr>
<td>dcat</td>
<td>DCAT 2 Vocabulary</td>
<td><a href="http://www.w3.org/ns/dcat#">http://www.w3.org/ns/dcat#</a></td>
</tr>
<tr>
<td>dct</td>
<td>DCMI Metadata Terms</td>
<td><a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a></td>
</tr>
<tr>
<td>skos</td>
<td>SKOS Simple Knowledge Organization System</td>
<td><a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a></td>
</tr>
<tr>
<td>xsd</td>
<td>XSD</td>
<td><a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a></td>
</tr>
<tr>
<td>prov</td>
<td>PROV</td>
<td><a href="http://www.w3.org/ns/prov#">http://www.w3.org/ns/prov#</a></td>
</tr>
<tr>
<td>foaf</td>
<td>FOAF</td>
<td><a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a></td>
</tr>
<tr>
<td>pav</td>
<td>PAV - Provenance, Authoring and Versioning</td>
<td><a href="http://pav-ontology.github.io/pav/">http://pav-ontology.github.io/pav/</a></td>
</tr>
</tbody>
</table>

Fig. 3. Key concepts and relationships.

where there is a complete overlap between the NE and GeoNames identifier.

Table 4 presents the statistics of NEs classes in .ann files (span-based) of the corpus.
5. Using the RDF version of LegalNERo

The LegalNERo corpus [34] is available for download from the Zeonodo platform\(^\text{14}\) as a single archive containing all the different representations described in this paper, stored into dedicated folders. In the “rdf” folder there is a single file containing all the triples in RDF-Turtle format. In addition to the download option, a SPARQL endpoint\(^\text{15}\) is available from the RELATE platform, hosted by the Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy.

The SPARQL endpoint is offered via an Apache Jena Fuseki server\(^\text{16}\). A simple graphical query interface, provided by the same server implementation, is available\(^\text{17}\). This allows a user to interact with the LegalNERo corpus by means of SPARQL queries and visualize results in table format. Figure 4 presents a SPARQL query to list legal references found in the corpus. It also shows the user interface displaying data in table form. This type of queries is useful in creating gazetteer resources specific to named entity recognition systems. The query can easily be adapted to produce lists of different types of entities.

Additional query examples are provided in Figures 5 and 6. In the first case, the SPARQL query allows listing of location entities with associated GeoNames identifiers. The result will contain only those entities that have a GeoNames identifier. Figure 6 makes use of the token layer available in the corpus and displays organization entities, tokenized, with the associated UPOS tags concatenated. In this example, only entities comprised of up to five tokens are considered. This type of query is useful in finding patterns associated with the named entities present in the corpus. Patterns can then be used with simpler pattern-based NER systems, such as Stanford RegexNER\(^\text{18}\), available from the Stanford CoreNLP\(^\text{25}\) package.

The advantage of having a linked data resource comes from the ability to interlink it with other resources. Recently, a number of other Romanian resources were converted to linked data format [2] and are available on the same SPARQL server as the LegalNERo corpus. This enables complex federated queries\(^\text{19}\) to be performed across multiple resources. An example is the refinement of the legal reference class into multiple fine-grained classes. Considering the classes "law", "decision", "government decision", and "government ordinances", they correspond to the presence in the entity of different forms of the Romanian words "lege", "decizie", "hotărâre", and "ordonanță". Thus, we can exploit the RoLEX lexicon to obtain the different word forms associated with the given words and then use these forms to classify the existing entities. An example query is given in Figure 7.

6. Corpus usage

In accordance with the multiple facets of the LegalNERo corpus, we developed two NER models [29]: one for all the entities and expressions, and another dealing only with persons, locations, organizations and time expressions. These models are based on a recurrent neural network with a final CRF layer, trained using the NeuroNER\(^\text{20}\) toolkit [9]. To improve the model’s performance, we used pre-trained word embeddings [33] representations trained on the Representative Corpus of Contemporary Romanian Language (CoRoLa) [43]. The models were integrated in the RELATE [28, 31] platform and are available for online interrogation and download\(^\text{21}\), together with the used word embeddings\(^\text{22}\).

In the context of the "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project\(^\text{23}\), we aim to develop an anonymization solution for Romanian language. Part of this solution, we need the identification of named entities present in a given document. Of course the purpose is not to anonymize leg-

---

\(^\text{14}\)https://doi.org/10.5281/zenodo.4772094
\(^\text{15}\)https://relate.racai.ro/datasets/legalner0/query
\(^\text{16}\)https://jena.apache.org/documentation/fuseki2/
\(^\text{17}\)https://relate.racai.ro/datasets/dataset.html?tab=query&ds=/legalner0
\(^\text{18}\)https://nlp.stanford.edu/software/regexner.html
\(^\text{19}\)https://www.w3.org/TR/sparql11-federated-query/
\(^\text{20}\)http://neuroner.com/
\(^\text{21}\)https://relate.racai.ro/index.php?path=ner/demo
\(^\text{22}\)http://relate.racai.ro/index.php?path=corola/we
\(^\text{23}\)https://curlicat-project.eu/
Table 3
NEs statistics on conllup files (token-based)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LEGAL</th>
<th>PER</th>
<th>LOC</th>
<th>ORG</th>
<th>TIME</th>
<th>GEO</th>
<th>TOTAL tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>conllup_PER_LOC_ORG_TIME</td>
<td>-</td>
<td>2,099</td>
<td>3,144</td>
<td>22,328</td>
<td>8,422</td>
<td>1,411</td>
<td>35,993</td>
</tr>
<tr>
<td>conllup_LEGAL_PER_LOC_ORG_TIME</td>
<td>24,687</td>
<td>2,099</td>
<td>3,144</td>
<td>19,477</td>
<td>5,121</td>
<td>1,411</td>
<td>54,528</td>
</tr>
</tbody>
</table>

Table 4
NEs statistics on .ann files (span-based)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LEGAL</th>
<th>PER</th>
<th>LOC</th>
<th>ORG</th>
<th>TIME</th>
<th>GEO</th>
<th>TOTAL NEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ann_PER_LOC_ORG_TIME</td>
<td>-</td>
<td>914</td>
<td>2,276</td>
<td>6,209</td>
<td>4,643</td>
<td>-</td>
<td>14,042</td>
</tr>
<tr>
<td>ann_LEGAL_PER_LOC_ORG_TIME</td>
<td>3,387</td>
<td>914</td>
<td>2,276</td>
<td>4,824</td>
<td>2,213</td>
<td>-</td>
<td>13,614</td>
</tr>
<tr>
<td>ann_LEGAL_PER_LOC_ORG_TIME_overlap</td>
<td>3,387</td>
<td>914</td>
<td>2,276</td>
<td>6,209</td>
<td>4,643</td>
<td>-</td>
<td>17,429</td>
</tr>
</tbody>
</table>

Fig. 4. SPARQL query to list legal references and corresponding result.

islation (which does not require anonymization), but we consider that the NER models developed based on the LegalNERo corpus, have the ability to complement other models developed on more general corpora and rule-based approaches. A current prototype of the anonymization solution is available through the RELATE platform and can be used online.\(^2\)

7. Conclusions and future work

This paper introduced the LegalNERo corpus. It is a manually annotated corpus for named entity recognition considering legal references in the Romanian language and is also enhanced with GeoNames identifiers. The corpus represents a subset of the larger MARCELL [44] comparable legislative corpus; therefore, for certain applications these corpora could be used together. LegalNERo also provides annotations for sub-entities present inside the legal references. This can

\(^2\)https://relate.racai.ro/index.php?path=roanon/anonymize
be exploited to allow usage of the corpus for training more classic NER systems considering only persons, locations, and organizations entities and time expressions.

We offer the corpus under a Creative Commons license (CC BY-ND 4.0). The downloadable version comes with different perspectives on the data, including span-based annotations, token-based annotations and RDF-Turtle format. We further offer a SPARQL endpoint allowing online interaction with the corpus. Finally, the corpus was integrated into the Linked Open Data Cloud.

Fig. 5. SPARQL query to list location entities with associated GeoNames identifiers.

```sql
PREFIX : <http://racai.ro/legalnero>
PREFIX powla: <http://purl.org/owls/powla/owl#>
PREFIX nerr: <http://nord.eurecom.fr/ontology#>
PREFIX gni: <http://www.geonames.org/ontology#>

SELECT ?id ?ent ?geo
WHERE {
  ?id a nerr:location .
  ?id powla:string ?ent .
}
LIMIT 25
```

Our aim is to further use this corpus to construct an improved NER system for the legal domain, in the Romanian language. Currently available models, presented in Section 6, achieved an average F1 score of 84% (considering all entities) and 84.70% (without the legal reference entity type). This already presents an improved performance compared to the one [30] previously used to automatically annotate the Romanian Legal Corpus [42] (part of the larger MARCELL corpus). We re-evaluated the old NER system ([30]) on the LegalNERo corpus and it achieved only a 49.38% average F1 score (with individual F1 scores 84.06% for time expressions, 56.7% for organizations, 26.85% for locations, and 19.3% for persons). This difference in performance comes from the lack of legal-domain text used in training the old NER system. Even though the size of the corpus is small, it has proven useful in improving the performance of the NER system and can be considered an important first resource for NER for the Romanian language in the legal domain. Nevertheless, considering additional techniques, such as word embeddings combinations [32] could prove beneficial in improving the overall performance.

Acknowledgements

Part of this work was conducted in the context of the “Curated Multilingual Language Resources for CEF.AT” (CURLICAT) project, CEF-TC-2019-1 – Automated Translation grant agreement number INEA/CEF/ICT/A2019/1926831.

References


Fig. 6. SPARQL query to list organization entities at token level (comprising up to five tokens) with associated UPOS tags.

Fig. 7. SPARQL query to classify legal reference entities into fine-grained classes.


Appendix A. Example encoding in the LegalNERo corpus

```sparql
:c1 a powla:Corpus, dcat:Dataset, prov:Entity ;
dct:title "LegalNERo" ;
dcat:theme :ner ;
dcat:distribution :legalnero.zip ;
```
nif:beginIndex "2" ;
nif:endIndex "36" .

dl_e2 a nerd:Time, powla:Node, nif:Phrase ;
powla:hasLayer :dl_l_ner ;
powla:string "17 octombrie 2019" ;
nif:anchorOf "17 octombrie 2019" ;
nif:beginIndex "19" ;
nif:endIndex "36" .