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

Vasile Păiș [a,*], Maria Mitrofan [a], Carol Luca Gasan [a], Alexandru Ianov [a], Corvin Ghiță [a], Vlad Silviu Coneschi [a], and Andrei Onuț [a]

[a] *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 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 [40], like persons, locations, organizations, time, proteins, etc. Starting with 1995, within the MUC-6 conference [14], 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 [36], named entities were considered as "phrases that contain the names of persons, organizations and locations". However, this limited approach is not suitable for every domain. In this context, in the biomedical domain, a number of works addressed entities such as genes, proteins, diseases [17], cell type [33], chemicals [13], [19].

In the legal domain, the TREC conference had a dedicated track [7] 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 [24]. The Competition on Legal Information Extraction and Entailment (COLIEE) [20] run over multiple editions allowed further exploration of tools and algorithms for information extraction in the legal domain.

In the context of the international project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL)[2] a large comparable corpus of legal documents for 7 languages was created [39]. This includes a monolingual sub-corpus for the Romanian language [37]. The Romanian corpus, as well as the other MARCELL corpora, was split at sentence and token level, lemmatized, and annotated at 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 these annotations were realized using automatic processes. Named entities were identified using a general-purpose tool [26], 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 time expressions. The tool was not trained on any legal texts.

Existing Romanian named entity corpora include: RONEC [10], Romanian TimeBank [11] and SiMoNERo [1]. The RONEC corpus contains 26,377 named entities, belonging to 16 different classes. The

*Corresponding author. E-mail: vasile@racai.ro.
[1] https://www.nist.gov/

[2] https://marcell-project.eu/

Romanian TimeBank corpus is an annotated parallel corpus for temporal information. This corpus contains 26,635 temporal named entities such as events, instances, signals, etc. SiMoNERo is a gold standard corpus for 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, time expressions and biomedical entities. Nevertheless, none of these corpora contains legal texts or legal entities.

This paper presents a manually annotated corpus, comprised 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 3 we present 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 we finally conclude in Section 7.

2. Related Work

One of the first papers to discuss named entity recognition in the legal domain is that of [9]. 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. [3] explore using the LKIF[3] ontology [16] further mapped to the YAGO[4] ontology [35] in order 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. [12] 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 [21], where references to legal norms are considered.

Leitner et al. [22] introduce a German legal named entity corpus comprising 7 coarse-grained classes which 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.

3. Annotation process

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

We considered 5 classes: person (PER), location (LOC), organization (ORG), time (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. In the case of organizations, they must have some formally established association. Typical examples are businesses, government units and political parties. Locations are defined on a geographical basis and include countries, cities and other geographical areas. References are introduced similar to [21] and the coarse-grained class of [22], without additional sub-classes. Thus, they are references to legal documents such as laws, ordinances, government decisions, etc.

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. 30 documents (out of the 100) were also

[3] https://github.com/RinkeHoekstra/lkif-core
[4] https://yago-knowledge.org/
[5] https://www.racai.ro/en/
[6] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf

shared with other two annotators. This aspect was hidden from the annotators during the process but allowed us to later compute inter-annotator agreement (IAA). 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 [25]. Actual annotation was handled using the BRAT [7] annotation tool [34], integrated into the RELATE platform. This allowed the annotators to view one document at the time, select the identified entity with the mouse and then associate an entity type to the selected text span.

After the annotation process ended, we were able to compute inter-annotator agreement between each pair of annotators, using Coehn's Kappa measure. This was accomplished at token level and lead 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 inclusion of indicative words in the entities (for example "orașul 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 it easy by highlighting entities found by multiple annotators.

Once all the common annotations were merged we re-computed Coehn'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.

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 via crawling from the public Romanian legislative portal[8]. As described in [39], the texts were extracted from the original HTML format and converted into TXT files. For the purposes of constructing the LegalNERo corpus, we selected a number of 370 documents of similar size, issued in the last two years (2020-2021). We also performed an initial check to make sure 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.

We used UDPipe[9] 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[10]. 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 [32]. This implies that each token has an additional annotation with the associated entity, prefixed with one of "B-" (for entity beginning) or "I-" (for a 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 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, location and time entities.

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 as well as allows analysis of the corpus using RDF queries and linking to external databases.

[7] https://brat.nlplab.org/index.html
[8] http://legislatie.just.ro/
[9] https://ufal.mff.cuni.cz/udpipe
[10] https://universaldependencies.org/format.html

Location entities can be resolved to places in the real world using geographical databases, such as GeoNames[11]. 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.

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 [5],[6]. 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 [15], by declaring it as a "nif:Word" element, linked to a "nif:Sentence".

We further investigated the POWLA [4] ontology. This was also used by [6] 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" which then 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 format).

For the named entity annotations, we employed the NERD ontology [30]. It was previously mentioned [31] 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. 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 which is defined as a work in a legislative corpus, which applies to acts that have been legally enacted (whether or not they are still in force).

The GeoNames database is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. According to the information available on the website[12], it contains over 25 million geographical names and consists of over 11 million unique features whereof 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.

Table 1 presents the vocabularies used in the corpus. The key concepts and relationships expressed in the dataset are visualized in Figure 1. 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 [2] for constructing the diagram shown in Figure 1.

The corpus is comprised of 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. Tokens are linked to the corresponding sentences and offer all the CoNLL-U Plus information, including word form, lemma, universal part-of-speech, language-specific part-of-speech, morphological features, dependency information, named entity type and GeoNames identifier.

The named entities document layer contains elements from the NERD and European Legislation Identifier ontologies. The elements also inherit from "nif:Phrase", thus specifying the beginning and end positions for the associated strings. Furthermore, the GeoNames feature identifier is specified when available for corresponding "nerd:Location" entities.

4.3. Statistics

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. In Table 2 are presented general corpus statistics. There are

[11] https://www.geonames.org/

[12] https://www.geonames.org/about.html

Table 1
Used vocabularies

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

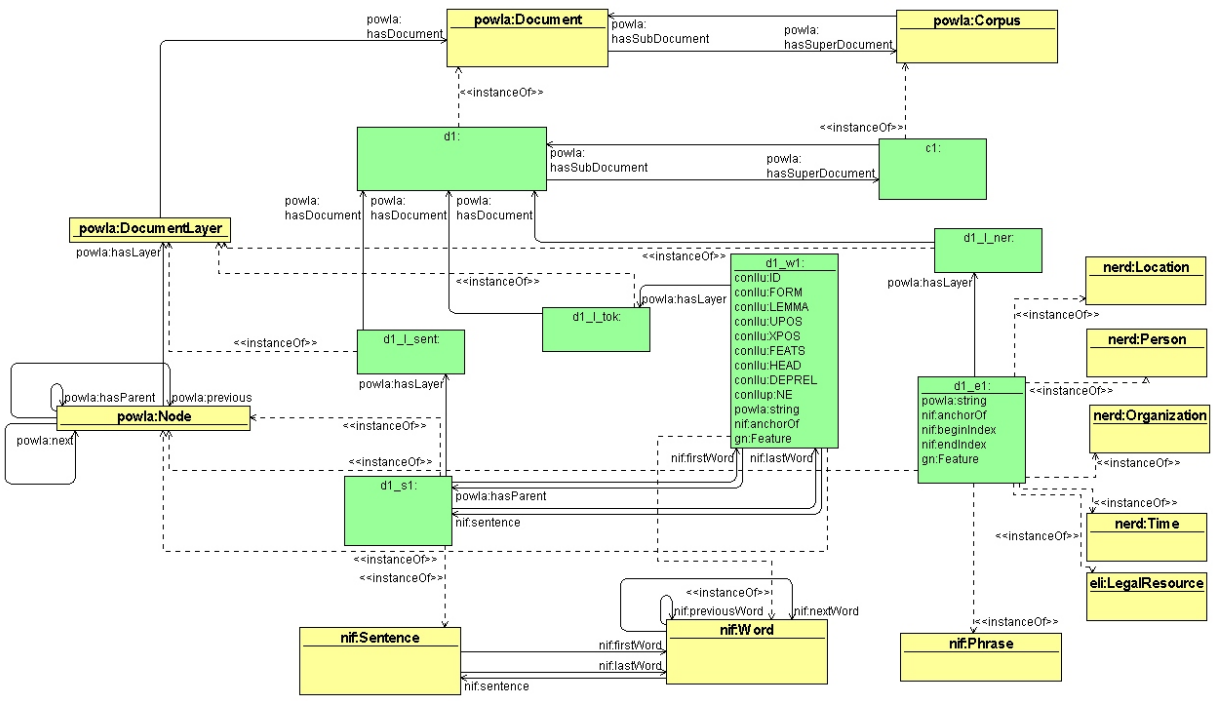

Fig. 1. Key concepts and relationships.

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 [18], 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 legal documents references class (LEGAL) contains

Table 2
Key statistics

Category	Value
Text Files	370
Tokens	265,335
Sentences	8,284
Unique lemma	12,887
Triples	5,761,781

2,851 organizations (ORG) and 3,301 time (TIME) NEs mentions. This format of the corpus also contains 1,411 GeoNames identifiers linked with the locations (LOC), where there is a complete overlapping between the NE and GeoNames identifier.

Table 4 presents the statistics of NEs classes in .ann files of the corpus.

5. Using the RDF version of LegalNERo

The LegalNERo corpus [29] is available for download from the Zeonodo platform[13] 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[14] 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[15]. A simple graphical query interface, provided by the same server implementation, is available[16]. This allows a user to interact with the LegalNERo corpus by means of SPARQL queries and visualize results in table format. Figure 2 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 3 and 4. 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 4 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 5 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[17], available from the Stanford CoreNLP [23] package.

6. Corpus usage

In accordance with the multiple facets of the LegalNERo corpus, we developed two NER models: one for all the entities and one dealing only with persons, locations, organizations and time entities. These models are based on a recurrent neural network with a final CRF layer, trained using the NeuroNER[18] toolkit [8]. To improve the model's performance, we used pre-trained word embeddings [28] representations trained on the Representative Corpus of Contemporary Romanian Language (CoRoLa) [38]. The models were integrated in the RELATE [25] platform and are available for online interrogation and download[19], together with the used word embeddings[20].

In the context of the "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project[21], 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 legislation (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[22].

[13] https://doi.org/10.5281/zenodo.4772094
[14] https://relate.racai.ro/datasets/legalnero/sparql
[15] https://jena.apache.org/documentation/fuseki2/
[16] https://relate.racai.ro/datasets/dataset.html?tab=query&ds=/legalnero
[17] https://nlp.stanford.edu/software/regexner.html
[18] http://neuroner.com/
[19] https://relate.racai.ro/index.php?path=ner/demo
[20] http://relate.racai.ro/index.php?path=corola/we
[21] https://curlicat-project.eu/
[22] https://relate.racai.ro/index.php?path=roanon/anonymize

Table 3
NEs statistics on conllup files (token-based)

Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL tokens
conllup_PER_LOC_ORG_TIME	-	2,099	3,144	22,328	8,422	1,411	35,993
conllup_LEGAL_PER_LOC_ORG_TIME	24,687	2,099	3,144	19,477	5,121	1,411	54,528

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

Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL NEs
ann_PER_LOC_ORG_TIME	-	914	2,276	6,209	4,643	-	14,042
ann_LEGALL_PER_LOC_ORG_TIME	3,387	914	2,276	4,824	2,213	-	13,614
ann_LEGAL_PER_LOC_ORG_TIME_overlap	3,387	914	2,276	6,209	4,643	-	17,429

```
prefix :        <http://racai.ro/legalnero>
prefix powla:   <http://purl.org/powla/powla.owl#>
prefix eli:     <http://data.europa.eu/eli/ontology#>

SELECT ?id ?ent
WHERE {
  ?id a eli:LegalResource .
  ?id powla:string ?ent .
}
LIMIT 5
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 5 of 5 entries

Search: Show 50 entries

	id	ent
1	:d338_e16	"Normelor metodologice de aplicare a Legii nr. 232 / 2016"
2	:d107_e7	"Referatul de aprobare al Direcției relații cu presa, afaceri europene și relații internaționale nr. S8 4.536 / 4.04.2019"
3	:d85_e20	"Legea nr. 13 / 2008"
4	:d291_e1	"ORDIN nr. 625 din 25 aprilie 2019"
5	:d319_e1	"ORDIN nr. 1.155 din 9 august 2019"

Showing 1 to 5 of 5 entries

Fig. 2. SPARQL query to list legal references and corresponding result.

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 also enhanced with GeoNames identifiers. The corpus represents a subset of the larger MARCELL [39] parallel legislative corpus, therefore for certain applications these corpora could be used together. LegalNERo provides also annotations for sub-entities present inside the legal references. This can be exploited to allow usage of the corpus for training more classic NER systems considering only persons, locations, organizations and time entities.

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.

```
prefix :        <http://racai.ro/legalnero>
prefix powla:   <http://purl.org/powla/powla.owl#>
prefix nerd:    <http://nerd.eurecom.fr/ontology#>
prefix gn:      <http://www.geonames.org/ontology#>

SELECT ?id ?ent ?geo
WHERE {
  ?id a nerd:Location .
  ?id powla:string ?ent .
  ?id gn:Feature ?geo .
}
LIMIT 25
```

Fig. 3. SPARQL query to list location entities with associated GeoNames identifiers.

Finally, the corpus was integrated in the Linked Open Data Cloud[23].

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 [26] previously used to automatically annotate the Romanian Legal Corpus [37] (part of the larger MARCELL corpus). Nevertheless, considering additional techniques, such as word embeddings combinations [27] 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

[1] Verginica Barbu Mititelu and Maria Mitrofan. The Romanian medical treebank-SiMoNERo. In *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2020*, pages 7–16, 2020.

[2] Jānis Bārzdiņš, Guntis Bārzdiņš, Kārlis Čerāns, Renārs Liepiņš, and Artūrs Sproģis. Uml style graphical notation and editor for owl 2. In Peter Forbrig and Horst Günther, editors, *Perspectives in Business Informatics Research*, pages 102–114, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16101-8.

[3] Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 9–18, 2017.

[4] Christian Chiarcos. Powla: Modeling linguistic corpora in owl/dl. In *Proceedings of the 9th International Conference on The Semantic Web: Research and Applications*, ESWC'12, pages 225–239, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642302831.

[5] Christian Chiarcos and Christian Fäth. Conll-rdf: Linked corpora done in an nlp-friendly way. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*, pages 74–88, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59888-8.

[6] Christian Chiarcos and Luis Glaser. A tree extension for CoNLL-RDF. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7161–7169, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.885.

[7] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2010 legal track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*, volume 500-294 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2010. URL https://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf.

[8] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2017.

[9] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. *Named Entity Recognition and Resolution in Legal Text*, pages 27–43. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12837-0.

[10] Ștefan Daniel Dumitrescu and Andrei-Marius Avram. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.546.

[11] Corina Forăscu and Dan Tufiş. Romanian timebank: An annotated parallel corpus for temporal information. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3762–3766, 2012.

[12] Ingo Glaser, Bernhard Waltl, and Florian Matthes. Named entity recognition, extraction, and linking in german legal contracts. In *IRIS: Internationales Rechtsinformatik Symposium*, pages 325–334, 2018.

[13] Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of*

[23] https://lod-cloud.net/dataset/racai-legalnero

```
prefix :        <http://racai.ro/legalnero>
prefix nif:     <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
prefix conllu:  <https://universaldependencies.org/format.html#>
prefix conllup: <https://universaldependencies.org/ext-format.html#>

SELECT ?id ?wp1 ?wp2 ?wp3 ?wp4 ?wp5
WHERE {
  ?id conllup:NE "B-ORG" .  ?id conllu:FORM ?word1 .  ?id conllu:UPOS ?pos1 .
  OPTIONAL { ?id nif:nextWord ?id2 . ?id2 conllup:NE "I-ORG" . ?id2 conllu:FORM ?word2 . ?id2 conllu:UPOS ?pos2 .
    OPTIONAL { ?id2 nif:nextWord ?id3 . ?id3 conllup:NE "I-ORG" . ?id3 conllu:FORM ?word3 .  ?id3 conllu:UPOS ?pos3 .
      OPTIONAL { ?id3 nif:nextWord ?id4 . ?id4 conllup:NE "I-ORG" . ?id4 conllu:FORM ?word4 . ?id4 conllu:UPOS ?pos4 .
        OPTIONAL { ?id4 nif:nextWord ?id5 . ?id5 conllup:NE "I-ORG" . ?id5 conllu:FORM ?word5 . ?id5 conllu:UPOS ?pos5 .
  } } } }

  BIND(CONCAT(STR(?word1),"/",STR(?pos1)) as ?wp1) .
  BIND(CONCAT(STR(?word2),"/",STR(?pos2)) as ?wp2) .
  BIND(CONCAT(STR(?word3),"/",STR(?pos3)) as ?wp3) .
  BIND(CONCAT(STR(?word4),"/",STR(?pos4)) as ?wp4) .
  BIND(CONCAT(STR(?word5),"/",STR(?pos5)) as ?wp5) .

}
```

Fig. 4. SPARQL query to list organization entities at token level (comprising up to 5 tokens) with associated UPOS tags.

The 5th Workshop on BioNLP Open Shared Tasks, pages 1–10, Hong Kong, China, November 2019. Association for Computational Linguistics. . URL https://www.aclweb.org/anthology/D19-5701.

[14] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL https://www.aclweb.org/anthology/C96-1079.

[15] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating nlp using linked data. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web – ISWC 2013*, pages 98–113, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41338-4.

[16] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. Lkif core: Principled ontology development for the legal domain. In *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 21–52, NLD, 2009. IOS Press. ISBN 9781586039424.

[17] Yuting Hu and Suzan Verberne. Named entity recognition for Chinese biomedical patents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 627–637, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. . URL https://www.aclweb.org/anthology/2020.coling-main.54.

[18] Radu Ion, Elena Irimia, Dan Ștefănescu, and Dan Tufiș. Rombac: The romanian balanced annotated corpus. In *LREC*, pages 339–344. Citeseer, 2012.

[19] Radu Ion, Vasile Păiș, and Maria Mitrofan. RACAI's system at PharmaCoNER 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 90–99, Hong Kong, China, November 2019. Association for Computational Linguistics. . URL https://www.aclweb.org/anthology/D19-5714.

[20] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In Kazuhiro Kojima, Maki Sakamoto, Koji Mineshima, and Ken Satoh, editors, *New Frontiers in Artificial Intelligence*, pages 177–192, Cham, 2019. Springer International Publishing. ISBN 978-3-030-31605-1.

[21] Jörg Landthaler, Bernhard Waltl, and Florian Matthes. Unveiling references in legal texts-implicit versus explicit network structures. In *IRIS: Internationales Rechtsinformatik Symposium*, volume 8, pages 71–8, 2016.

[22] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. Fine-grained named entity recognition in legal documents. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33220-4.

[23] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

[24] Douglas W. Oard, Jason R. Baron, Bruce Hedin, David D. Lewis, and Stephen Tomlinson. Evaluation of information retrieval for e-discovery. *Artif. Intell. Law*, 18(4):347–386, December 2010. ISSN 0924-8463. . URL https://doi.org/10.1007/s10506-010-9093-9.

[25] Vasile Păiș, Radu Ion, and Dan Tufiș. 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, May 2020. European Language Resources Association. ISBN 979-10-95546-64-1. URL https://www.aclweb.org/anthology/2020.iwltp-1.13.

[26] Vasile Păiș. *Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language*. PhD thesis, Romanian Academy, 2019.

[27] Vasile Păiș and Maria Mitrofan. Assessing multiple word embeddings for named entity recognition of professions and occupations in health-related social media. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 128–130, Mexico City, Mexico, June 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.smm4h-1.27.

[28] Vasile Păiș and Dan Tufiș. Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, 19(2):185–191, 2018.

[29] Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. Romanian Named Entity Recognition in the Legal domain (LegalNERo), May 2021. URL https://doi.org/10.5281/zenodo.4772094.

[30] Giuseppe Rizzo and Raphaël Troncy. NERD: A framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-2015.

[31] Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Brümmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW, 5th Workshop on Linked Data on the Web, April 16, 2012, Lyon, France*, Lyon, FRANCE, 04 2012. URL http://nerd.eurecom.fr/ui/paper/Rizzo_Troncy_Hellmann_Bruemmer-ldow2012.pdf.

[32] Erik F Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics, 1999.

[33] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 107–110, Geneva, Switzerland, August 28th and 29th 2004. COLING. URL https://www.aclweb.org/anthology/W04-1221.

[34] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-2021.

[35] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. . URL http://doi.acm.org/10.1145/1242572.1242667.

[36] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

[37] Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. Collection and annotation of the Romanian legal corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2773–2777, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.337.

[38] Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Onofrei Mihaela. Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary romanian. *Revue Roumaine de Linguistique*, 64(3):227–240, 2019.

[39] 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. The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.464.

[40] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1182.