Named entity recognition in the Romanian legal domain

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, Alexandru Ianov
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
vasile,maria@racai.ro

Abstract

Recognition of named entities present in text is an important step towards information extraction and natural language understanding. This work presents a named entity recognition system for the Romanian legal domain. The system makes use of the gold annotated LegalNERo corpus. Furthermore, the system combines multiple distributional representations of words, including word embeddings trained on a large legal domain corpus. All the resources, including the corpus, model and word embeddings are open sourced. Finally, the best system is available for direct usage in the RELATE platform.

1 Introduction

Natural language processing for the legal domain has its own unique challenges. This is due to the way legal documents are structured as well as to the domain-specific language being used. Technology dealing with legal documents has received increased attention in recent years. This can be seen from the number of recent scientific papers being published, the existence of the Natural Legal Language Processing (NLLP) workshop (Aletras et al., 2019, 2020) and different international projects dealing with natural language processing for the legal domain.

Named entity recognition (NER) is the process of identifying text spans that refer to real-world objects, such as organizations or persons, etc. One of the annotation schemes being used in a large number of works was introduced by the CoNLL-2003 shared task on language independent NER (Tjong Kim Sang and De Meulder, 2003) and refers to names of persons, organizations and locations. This annotation scheme can be applied in the legal domain as well, thus allowing existing systems to try to annotate legal documents (with or without being adapted to legal text). However, domain-specific entities are usually added to enhance the information extraction capabilities of text processing algorithms specifically designed for the legal domain. Dozier et al. (2010), while dealing with depositions, pleadings and trial-specific documents, propose including entities for attorneys, judges, courts and jurisdictions. Glaser et al. (2018) proposed adding the entities date, money value, reference and "other" for analyzing legal contracts. Leitner et al. (2019, 2020) proposed using 7 coarse-grained entity classes which can be further expanded into 19 fine-grained classes.

In the context of the "Multilingual Resources for CEF.AT in the legal domain" (MARCELL) project [1] a large, clean, validated domain-specific corpus was created. It contains monolingual corpora extracted from national legislation (laws, decrees, regulations, etc.) of the seven involved countries, including Romania (Tufiș et al., 2020). All seven corpora are aligned at topic level domains. The Romanian corpus was preprocessed (split at sentence level, tokenized, lemmatized and annotated with POS tags) using tools developed in the Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy (RACAI). Named entities were identified using a general-purpose tool (Păiș, 2019). This tool was designed for general Romanian language and allowed only four entity types: organizations, locations, persons and time expressions. The tool was not trained on any legal texts.

For the purposes of this work, we created a manually annotated corpus, comprising legal documents extracted from the larger MARCELL-RO corpus. We choose an annotation scheme covering 5 entity classes: person (PER), location (LOC), organization (ORG), time expressions (TIME) and legal document references (LEGAL). References are introduced similar to the work of Landthaler et al. (2016) and the coarse-grained class proposed by Leitner et al. (2019), without additional sub-classes.

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

Natural Legal Language Processing Workshop 2021, pages 9–18
November 10, 2021. ©2021 Association for Computational Linguistics

Thus, they are references to legal documents such as laws, ordinances, government decisions, etc. For the purposes of this work, in the Romanian legal domain, we decided to explore only these coarse-grained classes, without any fine-grained entities. This has the advantage of allowing the corpus to be used together with other general-purpose NER corpora. Furthermore, it allows us to judge the quality of the resulting NER system against existing systems. In order to train domain-specific NER systems, we constructed distributional representations of words (also known as word embeddings) based on the entire MARCELL corpus. Finally, we explored several neural architectures and adapted them as needed to the legal domain.

This paper is organized as follows: in Section 2 we present related work, in Section 3 is described the LegalNERo corpus, Section 4 presents the legal domain word embeddings, Section 5 describes the NER system architecture, while Section 6 gives the results and finally conclusions are presented in Section 7.

2 Related work

Legal NER is an important task in extracting key information from legal documents, such as dates, references to different types of legal documents, locations, organizations and persons. As Zhong and Tang (2020) stated, once the NEs are identified and classified they can be used in workflows to perform different functionalities such as document anonymization or case summarization.

One of the pioneering work in this research area was made by Dozier et al. (2010). The authors examined legal NEs in US depositions, pleadings, case law and other legal documents using statistical models, context rules, and a lookup list of NEs. They also developed different taggers for NEs such as document type of jurisdiction, obtaining an F1 score of 0.92 for the NEs belonging to jurisdiction class. This work formed the basis for Cardellino et al. (2017) to develop a tool for identifying, classifying and linking legal NEs. They trained and evaluated different systems (Stanford NER [2], a Support Vector Machine, and a neural network (NN)) on Wikipedia and on decisions coming from the European Court of Human Rights, obtaining an F1 score of 0.86 using NN and an F1 score of 0.56 using Stanford NER.

Glaser et al. (2018) studied and evaluated on German legal data three NER systems. GermaNER, a generic German NER tagger, has been adapted to identify and classify NEs such as persons, organizations, locations, and dates and money values. The second NER system used was DBpedia Spotlight pipeline, an interlinking hub, a tool that can be used to perform annotation tasks on a text provided by a user (Mendes et al., 2011). The third NER system employed in this task was based on contract templates to identify NEs (Minakov et al., 2007). For GermaNER pipeline and DBpedia Spotlight pipeline the evaluation was performed on a corpus of 500 judgements, and achieved an F1 score of 0.8 and 0.87 respectively. The template NER system was evaluated on a corpus of contract templates and obtained an F1 score of 0.92.

Leitner et al. (2019) also evaluated two systems, based on Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs), on NER for German language documents from the legal domain. The evaluation was performed on a German court decisions corpus annotated with 19 fine-grained classes, and also with 7 generalised coarse-grained classes. The best performance achieved by the CRFs was 93.23 on fined-grained classes and 93.22 on coarse-grained classes, and the BiLSTMs models reach a maximum of 95.46 for the fined-grained classes and 95.95 for coarse-grained ones. In the Lynx project[3] (Moreno-Schneider et al., 2020) a set of services, including NER, were developed in order to help with the creation of a legal domain knowledge graph (Legal Knowledge Graph – LKG) and its use for the semantic analysis of documents in the legal domain.

Barriere and Fouret (2019) described a method to generate contextual dictionaries for NER. The system was evaluated on a French legal corpus of 94 court decisions (276,705 tokens), which was annotated with 4 classes of entities. The best performance of this system was 96.52.

Even though there are several NER systems trained for Romanian language both for general language (Păiș, 2019) and specialised domains (Mitrofan, 2019), regarding the legal domain, the experiments are very few and the performances are low (for example Tufiș et al. (2020) note an average precision of 64.1% on a random sample extracted from the MARCELL-RO corpus, using the system devel-

[2] https://nlp.stanford.edu/software/CRF-NER.html

[3] https://www.lynx-project.eu/

oped in (Păiș, 2019)). Apart from the new LegalNERo corpus (see Section 3), existing Romanian NER corpora do not focus on the legal domain. Romanian TimeBank (Forăscu and Tufiş, 2012) is an annotated parallel corpus for temporal information. The RONEC (Dumitrescu and Avram, 2020) news corpus contains 26,377 named entities, belonging to 16 different classes. SiMoNERo (Barbu Mititelu and Mitrofan, 2020) is a gold standard corpus for biomedical domain, manually annotated with four types of domain-specific named entities.

3 The LegalNERo corpus

Annotation of the LegalNERo corpus was performed by 5 human annotators, supervised by two senior researchers at the Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (RACAI). For annotation purposes we used the BRAT tool[4] (Stenetorp et al., 2012), integrated in the RELATE platform (Păiș et al., 2020). Inside the legal reference class, we considered subentities of type organization and time. This allows for using the LegalNERo corpus in two scenarios: using all the 5 entity classes or using only the remaining general-purpose classes.

The LegalNERo corpus contains a total of 370 documents from the larger MARCELL-RO corpus. These documents were split amongst the 5 annotators, with certain documents being annotated by multiple annotators. Each annotator manually annotated 100 documents. The annotators were unaware of the overlap, which allowed us to compute an inter-annotator agreement. We used the Cohen's Kappa measure and obtained a value of 0.89, which we consider to be a good result.

The raw annotations were obtained in the BRAT specific format, consisting of text spans, characterized by start and end positions with the associated entities. However, since many NER systems make use of token-based annotations, we further employed the Romanian pipelines integrated in the RELATE platform (Păiș, 2020) and annotated the corpus at token level. For tokenization, lemmatization, part-of-speech tagging and dependency parsing we used UDPipe. Finally, the named entity annotations were mapped to tokens and exported in CoNLL-U Plus format[5], similar to the original format being used in the MARCELL-RO corpus.

Statistic	**Value**
Documents	370
Sentences	8,284
Tokens	265,335
Unique lemmas	12,887
RDF Triples	5,761,781

Table 1: LegalNERo corpus statistics

Entity	**Number**
Person	914
Location	2,276
Organization	4,824
Time	2,213
Legal Ref	3,387
Total	*13,614*

Table 2: Number of entities, considering all the entity types

Additionally, location entities were mapped to the GeoNames[6] ontology, but this information was not used for the purposes of this work. Nevertheless, the information is available in the LegalNERo corpus for future use. Finally, since we have multiple annotation levels available, we converted all the data into RDF format, specific to Linguistic Linked Open Data (LLOD) and made this available in the Linked Open Data Cloud[7].

Key statistics computed on the LegalNERo corpus are presented in Table 1. The number of entities, considering all the entity types, are given in Table 2, while considering only persons, locations, organization and time expressions are given in Table 3. These numbers are obtained at entity level (not at token level). We further computed in Table 4 the average number of tokens associated with each entity type.

Results presented in Table 4 clearly indicate that the legal reference entity type has the largest number of tokens (7.29 in average). A typical example is "ORDIN nr. 625 din 25 aprilie 2019" ("Order no. 625 from 25 April 2019"). Nevertheless, longer entities are also present, such as "Ordinul președintelui Casei Naționale de Asigurări de Sănătate nr. 141 / 2017" ("Order of the president of the National Health Insurance House no. 141 / 2017"). This example has 12 tokens and contains an orga-

[4] https://brat.nlplab.org/

[5] https://universaldependencies.org/ext-format.html

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

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

Entity	Number
Person	914
Location	2,276
Organization	6,209
Time	4,643
Total	*14,042*

Table 3: Number of entities, without the legal reference entity type

Entity	# Tokens
Person	2.30
Location	1.38
Organization	4.04
Time	2.31
Legal Ref	7.29

Table 4: Average number of tokens for each entity type

nization sub-entity ("Casei Naționale de Asigurări de Sănătate" / "National Health Insurance House") as well as a time expression ("2017").

4 Romanian legal-domain word embeddings

Previously computed Romanian word embeddings (Păiș and Tufiș, 2018) made use of the Representative Corpus of Contemporary Romanian Language (CoRoLa) (Tufiș et al., 2019). This large corpus contains texts from multiple domains, including the legal domain. However, in addition to the legal domain, the CoRoLa corpus contains texts from completely different areas, such as news, literature and mathematics. This makes the CoRoLa-based embeddings suitable for general tasks, but also allows for creation of legal-domain specific representations.

For the purposes of this work, in addition to the already available CoRoLa embeddings, we constructed word representations based on the entire MARCELL-RO corpus. We used the same approach which obtained the best performing CoRoLa embeddings. Thus, we used the FastText[8] toolkit (Joulin et al., 2017) and produced vector representations of dimension 300, while considering only words appearing a minimum of 20 times. Furthermore, the model made use of n-gram windows of dimension 5. The resulting embeddings are available for download within the RELATE platform[9].

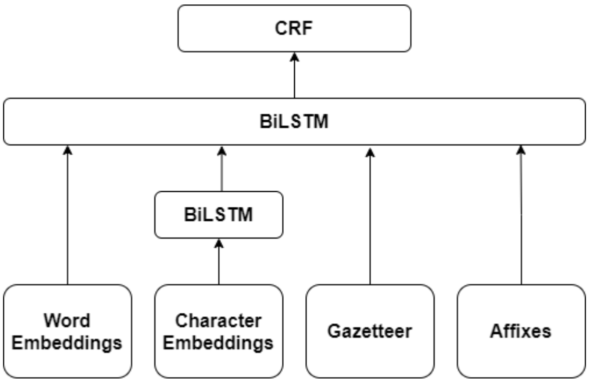


Figure 1: System architecture

For Romanian language there are currently no legal domain contextual embeddings, like the LegalBERT (Chalkidis et al., 2020) model for English. Furthermore, existing Romanian BERT models, like (Masala et al., 2020; Ștefan Daniel Dumitrescu et al., 2020), were not trained on corpora containing legal documents.

5 System architecture

Our proposed NER model makes use of recurrent neural networks, based on BiLSTM cells, with a final CRF layer. For features we considered word representations, character embeddings, gazetteer resources and known affixes. The word embeddings are initialized from a pre-trained model and are fine-tuned during training. The character embeddings are computed during training and the embedding layer is followed by a BiLSTM layer helping with the representation generation. For implementation we used a modified version (Armengol-Estapé et al., 2019) of the NeuroNER (Dernoncourt et al., 2017) software package. This implementation was further adapted to our needs in order to allow online model serving. The overall system diagram is presented in Figure 1.

To construct the gazetteer resources we employed the GeoNames database for the country Romania and the JRC-Names[10] (Steinberger et al., 2011) multilingual named entity collection. These two collections cover a large number of entity names like locations, organizations and persons.

We trained multiple models using different features. This includes different word embeddings

[8] https://fasttext.cc/

[9] https://relate.racai.ro/index.php?path=ner/demo

[10] https://ec.europa.eu/jrc/en/language-technologies/jrc-names

(CoRoLa and MARCELL embeddings) and also combinations of the two embeddings. Previous work (Păiș and Mitrofan, 2021), (Casillas et al., 2019) has shown that using different word embeddings and combinations can improve NER performance. For each word representation we adapted the main BiLSTM layer size to match the embedding size. For example, in the case of CoRoLa embeddings we used a layer with size 300 and in the case of CoRoLa + MARCELL embeddings we used a layer with size 600. This was done to accommodate the increased vector size associated with the word representation.

In addition to the main BiLSTM layer size, we used a character BiLSTM of size 25. Furthermore, to prevent overfitting, a dropout of probability 50% was introduced. A gradient clipping (Pascanu et al., 2013) with a value of 5 is used to deal with exploding gradients. Finally, we use a stochastic gradient descent (SGD) algorithm with a learning rate of 0.005.

Besides experimenting with the aforementioned NER models, we also made tests with an ensemble method that combines the results from the different models. The idea was to see what combination provides better results and in what conditions. We've used four types of operations for combining the results. The first one is the union of two or more models. The second one is the full intersection of two or more models. In this case, if for the same set of tokens the annotations are different then none of them are represented in the final set. The third type is the implementation of a majority voting system, meaning that if an entity span has the same annotation in $(n/2) + 1$ cases, where n is the total number of models used, it is the winner. For example, given at least three models it takes at least two of the candidates with the same annotation to allow it to be represented in the final set. The fourth one is a merge between two or more models where for a given set of tokens the longest annotation between all of them is represented in the final set.

Afterwards, for each of the resulting final sets, the precision, recall and F1 scores are computed against the gold corpus for each entity types in the corpus. The macro average score is finally calculated. Each of these pairs of combinations and scores are then recorded and a best that has the highest F1 score is calculated. The final step of measuring the precision, recall and F1 scores was done using NeuroNER evaluation script, in order to have a consistency with the initial models scores.

6 Results

Experiments were performed on the two perspectives associated with the corpus: the complete set of 5 classes (person, organization, location, time expressions and legal reference) and the reduced set (person, organization, location, time expression). In the second case, we took into consideration the additional annotations associated with the 4 remaining classes present inside the legal reference entities. Furthermore, in order to explore the impact of different features, we conducted multiple experiments on each perspective. Each experiment was allowed to train for at most 100 epochs. However, we used early stopping in case no improvement was perceived on the validation set for 10 epochs. As a result, neither of the experiments actually trained for the maximum of 100 epochs.

The LegalNERo corpus was split into three sub-corpora for training, validation during training and testing on data unseen during training. We used a 80% split for training and 10% for each of the validation and testing sub-corpora. The split was realized at file level, thus the training split contains 290 files, while the validation and test splits contain 40 files each. For reproducibility of the reported results as well as for comparison with future models, we offer the splits for download from the RELATE platform[11],[12].

Using the same splits, we also trained baseline models using well-known libraries such as spaCy [13]. The spaCy library provides a variety of tools for fast text processing and is developed as a modular pipeline. For implementation of our models we focused onto two critical components of the spaCy pipeline, namely the components which are responsible for converting string tokens to vectors and the named entity recognition component. During training, besides word embeddings, we used spaCy's class Lexeme [14], as an entry in the vocabulary. A Lexeme has no string context, it is a word type, as opposed to a word token. It therefore has no part-of-speech tag, dependency parse, or lemma (if lemmatization depends on the part-of-speech tag). After exhaustive tests, the representative model for each

[11] https://relate.racai.ro/resources/legalnero/legalnero_split_5classes.zip

[12] https://relate.racai.ro/resources/legalnero/legalnero_split_4classes.zip

[13] https://spacy.io/

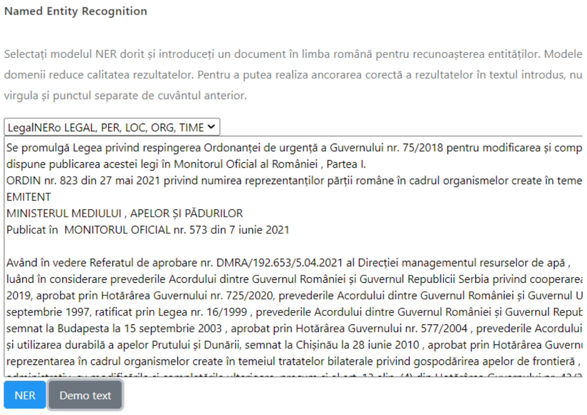
[14] https://spacy.io/api/lexeme

word embedding source also had its own optimal selection of the parameters that can be found at [15]. All of them made use of at least one of the affixes and some of them also used the shape. The spaCy's training process, which is stopped automatically when the F1 score doesn't vary very much for a couple of epochs, guarantees saving the best model from all checkpoints. Although it doesn't save any information during training, we used wandb [16] in order to retrieve the aforementioned data. Using spaCy, the best F1 score was 83.77, for all the five entity classes and using ro_core_news_lg[17] embeddings with a layer of size 300. When the LEGAL class was excluded and with the same embeddings configuration, the F1 score obtained by spaCy is 86.21.

In Table 5 are presented different experiments using the system architecture described in Section 5, for all the entities. Similarly, in Table 6 are given results from experiments without the legal reference entity class. The usage of gazetteer resources helps to increase the F1 score associated with persons, locations and organizations. Therefore the best models associated with the two scenarios make use of gazetteers.

Table 7 presents the results of the four types of ensemble operations. It can be seen that the best F1 score (90.36) was achieved by reunion of three models, two containing all types of entities and one obtained without legal entity type (CoRoLa+MARCELL Y N). The model, CoRoLa+MARCELL Y N presented in Table 3 has significantly contributed to increasing the F1 score. Another important observation presented in Table 7 is that an ensemble of models can, in principle, perform better than any individual model, because the various errors of the models were averaged out. It can also be seen that the F1 score for each operation is greater than the ones obtained by individual models.

The resulting best performing models are available for direct online usage through the RELATE platform[18]. This integration allows the user to enter a Romanian legal document inside the platform's web interface, select the desired model, by using a dropdown menu, and then execute the model. The selected model together with the raw text are sent to the server process, which produces a list of recognized entities. These are returned to the user and displayed in the web interface. The interface is presented in Figure 2 and example results are presented in Figure 3. Furthermore, pre-trained models can be downloaded from the same interface.


Named Entity Recognition

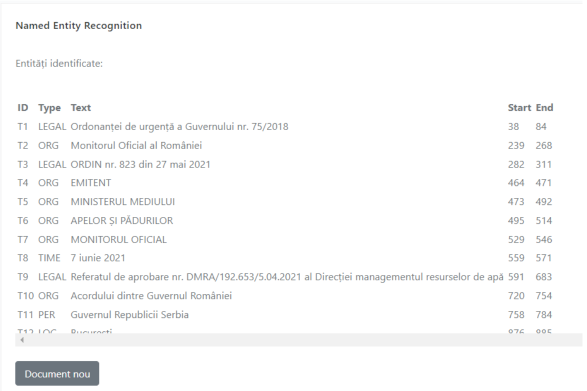
Selectați modelul NER dorit și introduceți un document în limba română pentru recunoașterea entităților. Modele
domenii reduce calitatea rezultatelor. Pentru a putea realiza ancorarea corectă a rezultatelor în textul introdus, nu
virgula și punctul separate de cuvântul anterior.

LegalNERo LEGAL, PER, LOC, ORG, TIME

Se promulgă Legea privind respingerea Ordonanței de urgență a Guvernului nr. 75/2018 pentru modificarea și comp
dispune publicarea acestei legi în Monitorul Oficial al României , Partea I.
ORDIN nr. 823 din 27 mai 2021 privind numirea reprezentanților părții române în cadrul organismelor create în teme
EMITENT
MINISTERUL MEDIULUI , APELOR ȘI PĂDURILOR
Publicat în MONITORUL OFICIAL nr. 573 din 7 iunie 2021

Având în vedere Referatul de aprobare nr. DMRA/192.653/5.04.2021 al Direcției managementul resurselor de apă ,
luând în considerare prevederile Acordului dintre Guvernul României și Guvernul Republicii Serbia privind cooperarea
2019, aprobat prin Hotărârea Guvernului nr. 725/2020, prevederile Acordului dintre Guvernul României și Guvernul U
septembrie 1997, ratificat prin Legea nr. 16/1999 , prevederile Acordului dintre Guvernul României și Guvernul Repub
semnat la Budapesta la 15 septembrie 2003 , aprobat prin Hotărârea Guvernului nr. 577/2004 , prevederile Acordului
și utilizarea durabilă a apelor Prutului și Dunării, semnat la Chișinău la 28 iunie 2010 , aprobat prin Hotărârea Guvern
reprezentarea în cadrul organismelor create în temeiul tratatelor bilaterale privind gospodărirea apelor de frontieră ,

Figure 2: Web interface for interacting with the Romanian Legal NER models


Named Entity Recognition

Entități identificate:

ID	Type	Text	Start	End
T1	LEGAL	Ordonanței de urgență a Guvernului nr. 75/2018	38	84
T2	ORG	Monitorul Oficial al României	239	268
T3	LEGAL	ORDIN nr. 823 din 27 mai 2021	282	311
T4	ORG	EMITENT	464	471
T5	ORG	MINISTERUL MEDIULUI	473	492
T6	ORG	APELOR ȘI PĂDURILOR	495	514
T7	ORG	MONITORUL OFICIAL	529	546
T8	TIME	7 iunie 2021	559	571
T9	LEGAL	Referatul de aprobare nr. DMRA/192.653/5.04.2021 al Direcției managementul resurselor de apă	591	683
T10	ORG	Acordului dintre Guvernul României	720	754
T11	PER	Guvernul Republicii Serbia	758	784
T12	LOC	București	876	885

Figure 3: Romanian Legal NER results presented in RELATE

7 Conclusions

This paper introduced a neural named entity recognition system designed specifically for the Romanian legal domain. It employed the LegalNERo corpus for training and evaluation. The system is available for querying inside the RELATE platform and pre-trained models are available for download. As indicated in Section 6, the best performing models made use of word embeddings trained on the legal-domain MARCELL corpus. When considering all the entity types available, CoRoLa and

[15] https://spacy.io/api/lexeme#attributes

[16] https://wandb.ai/site

[17] https://spacy.io/models/ro#ro_core_news_lg

[18] https://relate.racai.ro/index.php?path=ner/demo

Embeddings	Gaz.	Affixes	LEGAL	PER	LOC	ORG	TIME	Macro AVG
Validation set								
CoRoLa	N	N	84.58	81.75	76.11	80.07	84.12	81.37
CoRoLa	Y	N	84.88	80.67	76.73	80.11	83.72	81.26
CoRoLa	Y	Y	83.72	**83.00**	74.10	80.15	84.21	81.09
MARCELL	N	N	85.79	82.87	75.15	82.56	79.91	81.29
MARCELL	Y	N	82.75	78.88	77.44	**82.95**	**85.65**	**81.56**
MARCELL	Y	Y	**86.12**	81.30	73.58	81.10	82.48	80.97
CoRoLa+MARCELL	N	N	84.51	77.42	74.78	80.54	84.30	80.32
CoRoLa+MARCELL	Y	N	85.61	79.52	71.78	80.86	83.76	80.33
CoRoLa+MARCELL	Y	Y	83.84	77.11	**77.58**	80.78	81.78	80.24
Test set								
CoRoLa	N	N	**90.50**	95.56	70.59	76.26	**85.93**	83.90
CoRoLa	Y	N	90.06	98.08	75.37	78.38	82.42	85.03
CoRoLa	Y	Y	89.80	95.56	73.33	75.80	84.53	83.94
MARCELL	N	N	90.41	97.38	70.30	76.70	81.64	83.49
MARCELL	Y	N	86.98	98.48	**75.94**	**80.60**	84.09	**85.34**
MARCELL	Y	Y	90.12	96.65	69.77	74.23	85.55	83.39
CoRoLa+MARCELL	N	N	88.18	**98.50**	75.62	76.65	84.39	84.74
CoRoLa+MARCELL	Y	N	89.68	97.04	75.21	78.69	83.08	84.83
CoRoLa+MARCELL	Y	Y	89.42	96.99	70.03	79.10	80.54	83.40

Table 5: F1 scores for different models, considering all entities

Embeddings	Gaz.	Affixes	PER	LOC	ORG	TIME	Macro AVG
Validation set							
CoRoLa	N	N	80.50	73.65	**87.17**	82.68	81.17
CoRoLa	Y	N	80.63	**78.92**	85.96	83.07	**82.21**
CoRoLa	Y	Y	79.01	75.00	85.81	84.15	81.18
MARCELL	N	N	**81.75**	73.58	86.17	83.51	81.57
MARCELL	Y	N	79.35	75.53	85.47	**85.45**	81.78
MARCELL	Y	Y	80.65	71.97	86.40	83.94	81.10
CoRoLa+MARCELL	N	N	77.05	75.76	85.89	84.59	81.04
CoRoLa+MARCELL	Y	N	81.12	73.23	85.48	83.69	81.13
CoRoLa+MARCELL	Y	Y	76.54	75.29	85.99	85.12	81.05
Test set							
CoRoLa	N	N	96.27	66.86	80.34	89.81	83.35
CoRoLa	Y	N	96.65	72.13	81.36	88.31	84.64
CoRoLa	Y	Y	97.69	69.54	80.09	89.06	84.10
MARCELL	N	N	96.68	74.01	81.24	91.65	85.94
MARCELL	Y	N	**98.86**	69.83	79.85	91.93	85.14
MARCELL	Y	Y	96.68	74.49	78.87	**92.18**	85.66
CoRoLa+MARCELL	N	N	**98.86**	69.59	**82.13**	90.51	85.29
CoRoLa+MARCELL	Y	N	97.40	72.88	80.89	90.28	85.39
CoRoLa+MARCELL	Y	Y	**98.86**	**76.01**	80.89	91.39	**86.84**

Table 6: F1 scores for different models, considering only person, location, organization and time expression

Operation	LEGAL	PER	LOC	ORG	TIME	Macro AVG
Reunion	90.42	98.17	78.96	91.61	92.44	**90.36**
Intersection	90.12	98.66	65.45	87.67	86.64	86.14
Voting algorithm	90.41	99.33	75.05	91.18	90.57	89.37
Longest span	89.35	98.66	70.59	89.95	86.22	87.29

Table 7: F1 scores for different ensembles for test set, considering all entities

MARCELL embeddings seem to provide similar performance (Table 5, F1 difference on the test set is less than 1%). However, when considering only persons, organizations, locations and time expressions (Table 6), MARCELL embeddings provide over 1% F1 improvement compared to CoRoLa, while a combination of CoRoLa and MARCELL embeddings provide the best results with an improvement of over 2% over simple CoRoLa based embeddings. Even more, an ensemble model combining models with all the entity types with a model without the legal reference entity type achieves the best performance on the test set with almost 5% improvement, considering overall macro F1. As future work we foresee expanding the LegalNERo corpus with additional annotations including fine-grained classes of entities and to make more experiments with different NER architectures.

References

Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota.

Nikolaos Aletras, Androutsopoulos Ion, Leslie Barrett, Adam Meyers, and Daniel Preotiuc-Pietro, editors. 2020. *Proceedings of the Natural Legal Language Processing Workshop 2020*.

Jordi Armengol-Estapé, Felipe Soares, Montserrat Marimon, and Martin Krallinger. 2019. Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics Inform*, 17(2):e15–.

Verginica Barbu Mititelu and Maria Mitrofan. 2020. 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.

Valentin Barriere and Amaury Fouret. 2019. May i check again?–a simple but efficient way to generate and use contextual dictionaries for named entity recognition. application to french legal texts. *arXiv preprint arXiv:1909.03453*.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. 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.

Arantza Casillas, Nerea Ezeiza, Iakes Goenaga, Alicia Pérez, and Xabier Soto. 2019. Measuring the effect of different types of unsupervised word representations on medical named entity recognition. *International Journal of Medical Informatics*, 129:100–106.

Ștefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4324–4328. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

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

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named Entity Recognition and Resolution in Legal Text*, pages 27–43. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ștefan Daniel Dumitrescu and Andrei-Marius Avram. 2020. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France. European Language Resources Association.

Corina Forăscu and Dan Tufiş. 2012. 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.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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

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 12th Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Mihai Masala, Ștefan Ruseti, and Mihai Dascalu. 2020. RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8.

Igor Minakov, George Rzevski, Petr Skobelev, and Simon Volman. 2007. Creating contract templates for car insurance using multi-agent based text understanding and clustering. In *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*, pages 361–370. Springer.

Maria Mitrofan. 2019. *Extragere de cunostinte din texte în limba româna si date structurate cu aplicatii în domeniul medical*. Ph.D. thesis, Romanian Academy.

Julian Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodriguez-Doncel, Artem Revenko, Sotirios Karampatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia, and Filippo Maganza. 2020. Orchestrating NLP services for the legal domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2332–2340, Marseille, France. European Language Resources Association.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages 1310–1318. JMLR.org.

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

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 and Maria Mitrofan. 2021. 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. Association for Computational Linguistics.

Vasile Păiș and Dan Tufiș. 2018. 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.

Vasile Păiș, Dan Tufiș, and Radu Ion. 2020. A processing platform relating data and tools for romanian language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 81–88, Marseille, France. European Language Resources Association.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. 2011. JRC-NAMES: A freely available, highly multilingual named entity resource. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 104–110, Hissar, Bulgaria. Association for Computational Linguistics.

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.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. 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.

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 2773–2777, Marseille, France. European Language Resources Association.

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

Qing Zhong and Yan Tang. 2020. An attention-based bilstm-crf for chinese named entity recognition. In *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 550–555. IEEE.