Improving Romanian BioNER Using a Biologically Inspired System

Maria Mitrofan
RACAI, Romanian Academy
maria@racai.ro

Vasile Păiș
RACAI, Romanian Academy
vasile@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 biomedical domain. The system makes use of a new and extended version of SiMoNERo corpus, that is open sourced. Additionally, the best system is available for direct usage in the RELATE platform.

1 Introduction

The rapid advancement of Artificial Intelligence (AI) technologies has led to the development of different prominent fields of AI such as natural language processing (NLP). NLP is able to provide valuable information from large amounts of texts. For example, in the COVID-19 pandemic situation, NLP has played an important role in finding the presence of disease (Cury et al., 2021).

Identifying text spans that refer to real-world objects and categorizing them into a subject under an entity, is known as Named Entity Recognition (NER) (Nadeau and Sekine, 2007; Ananiadou et al., 2004). However, each domain has its own types of entities, for example, NER in the biomedical domain implies identifying chemicals, symptoms, ingredients, diseases, genes, dosage level, dosage forms, active substances, etc.

Although the NLP community has invested a lot of effort in developing BioNER systems for the English language, obtaining important results, the development of NER systems for other languages is conditioned by the availability of quality resources, such as gold annotated NER corpora. Moreover, biomedical NER has multiple specificities that one needs to address when developing an NER system: spelling variations, huge amounts of abbreviations, lengthy phrases, polysemy, cascaded constructions (Mitrofan, 2017). Consequently BioNER is a challenging task and most of the time NER systems need domain adaptation. In this paper we propose a NER system that uses pre-trained contextual embeddings, XLM-RoBERTa (Conneau et al., 2020), enhanced with an inhibitory mechanism similar to the biological process of lateral inhibition (Cohen, 2011), that has as the main goal the filtration of noisy information, in our case noise can be associated with rare contexts or less occurring entities. This system is trained on a new version of SiMoNERo corpus, whose NER level has been expanded with new entities (including COVID pandemic-related entities) for a better coverage of biomedical language.

This paper is organized as follows: in Section 2 we present related work, in Section 3 the SiMoNERo corpus is presented, Section 4 describes the NER system architecture, while Section 5 gives the results and finally conclusions are presented in Section 6.

2 Related work

BioNER is an important task that aims to extract key information from biomedical documents that can be used in workflows to perform different functionalities such as relation extraction, text mining, etc. In recent years, pre-trained models, such as BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2020) have made significant contributions to the development of the NER task.

In the context of the 2020 Iberian Languages Evaluation Forum (IberLEF) shared task, Xiong et al. (2020) used BERT as the base module and a machine reading comprehension framework was proposed to identify and classify NEs that achieved an F1-score of 0.87. Weber et al. (2021) developed HunFlair, a NER tagger, able to recognize five biomedical entity types. It outperforms other NER tagers with an average gain of 7.26% when compared with other state-of-the-art biomedical NER tools such as SciSpacy (Neumann et al., 2019) or HUNER (Weber et al., 2020). HunFlair uses a

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 316–322

character-level model that was pretrained on 3 million full texts and 24 million biomedical abstracts.

Even though the performance of NER systems for the biomedical domain for English has increased lately, there is still room for improvement until human annotators performance is reached. The Romanian language, suffers from the scarcity of NER systems for different subdomains, especially in the biomedical domain. One of the first attempts to develop a biomedical NER tagger was based on a Partitioned Convolutional Neural Network for classification and used word-embeddings computed from the Romanian section of Wikipedia, concatenated with a medical sub-corpus (Mitrofan, 2017). This approach achieved an F1-score of around 0.5. A more recent approach was based on Bidirectional Long-Short-Term Memory (BiLSTM) networks and obtained an F1-score of 0.81 (Mitrofan, 2019).

3 SiMoNERo corpus

SiMoNERo is the gold standard morphologically, syntactically and named entity annotated Romanian medical corpus. This corpus has three different development stages. The first one was the creation of the MoNERo corpus, a gold standard biomedical corpus for Romanian language enhanced with two types of annotations: morphological and named entities specific to the biomedical domain (Mitrofan et al., 2019). The second development stage was the addition of the syntactic annotations (Mititelu and Mitrofan, 2020). The current phase is the one in which the named entity annotation level was enhanced by 10%, due to the addition of new relevant sentences. Currently, SiMoNERo has 163,707 tokens, comprised in 5,418 sentences and 15,493 NEs.

SiMoNERo contains texts from three types of documents: scientific medical literature books, medical journal articles, and sites that offer explanations on various medical topics. Regarding the medical domain, the texts were chosen to belong mainly to cardiology, diabetes, and endocrinology.

The annotation scheme of the corpus has three different levels:

- The morphological level that had two development stages: automatic annotation using the TTL tool (Ion, 2007) and manual verification of each tag. Currently, the POS-tags of the newly added sentences are yet to be validated by hand.

Type	Average	Stdev.
ANAT	1.64	0.82
CHEM	1.34	0.73
DISO	1.78	0.99
PROC	1.85	0.99

Table 1: The average size of NEs

- Named entity level that was developed by two annotators. The annotation scheme contains four semantic groups: anatomy (ANAT), chemicals and drugs (CHEM), disorders (DISO), and procedures (PROC). Each entity is marked in Inside-Outside-Beginning (IOB2) format (Sang and Veenstra, 1999), where "B" denotes the beginning of a chunk (a span of tokens) and "I" represents an inside of a chunk. "O" - labels highlight tokens that do not belong to a chunk. Figure 2 presents an excerpt of the corpus with annotations ("Eritemul diabetic deseori mimează erizipelul și de aceea este numit și eritem pseudo-erizipeloid"/ "Diabetic erythema often mimics erysipelas and and therefore it is also called erysipeloid erythema"). In order to see the guidelines for named entity annotation see (Mitrofan et al., 2019). Currently, this level of annotation was expanded with 2,176 new NEs annotations: 385 (ANAT), 213 (CHEM), 566 (DISO), and 1,012 (PROC).

- Syntactic level that was added automatically using NLP-Cube parser (Boroş et al., 2018). Additionally, a manual validation process was performed to ensure compatibility with Universal Dependencies (UD)[1] validation tests.

After the corpus was expanded with new annotations regarding the named entities level, all sentences were shuffled and split into three files: train, dev, and test. In order to evaluate our approach we used 80% of the corpus sentences for training, 10% for development, and 10% for testing. Figure 2 shows the label distribution in the train, dev and test sets. Y axis indicates the number of a particular label in the data and Table 1 indicates that most of the medical NEs are compound of more than one token. This version of the corpus is freely available for download and non-commercial use [2].

[1] https://universaldependencies.org/
[2] https://www.racai.ro/tools/text/

```
# sent_id = test_46
# text = Eritemul diabetic deseori mimează erizipelul și de aceea este denumit și eritem pseudo-erizipeloid.
1   Eritemul    eritem  NOUN    Ncmsry  Case=Nom|Definite=Def|Gender=Masc|Number=Sing   4   nsubj   _   _   B-DISO
2   diabetic    diabetic    ADJ Afpms-n Definite=Ind|Degree=Pos|Gender=Masc|Number=Sing 1   amod    _   _   I-DISO
3   deseori deseori ADV Rqp Degree=Pos  4   advmod  _   _   O
4   mimează mima    VERB    Vmip3   Mood=Ind|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _   O
5   erizipelul  erizipel    NOUN    Ncmsry  Case=Nom|Definite=Def|Gender=Masc|Number=Sing   4   obj _   _   B-DISO
6   și  și  CCONJ   Crssp   Polarity=Pos    10  cc  _   _   O
7   de  de  ADP Spsa    AdpType=Prep|Case=Acc   10  advmod  _   _   O
8   aceea   acela   PRON    Pd3fsr  Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Dem   7   fixed   _   _   O
9   este    fi  AUX Vaip3s  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   10  aux:pass    _   _   O
10  denumit denumi  VERB    Vmp--sm Gender=Masc|Number=Sing|VerbForm=Part   4   conj    _   _   O
11  și  și  CCONJ   Crssp   Polarity=Pos    12  cc  _   _   O
12  eritem  eritem  NOUN    Ncms-n  Definite=Ind|Gender=Masc|Number=Sing    10  conj    _   _   B-DISO
13  pseudo-erizipeloid  pseudo-erizipeloid  ADJ Afpms-n Definite=Ind|Degree=Pos|Gender=Masc|Number=Sing 12  advmod  _   SpaceAfter=No   I-DISO
14  .   .   PUNCT   PERIOD  _   4   punct   _   SpacesAfter=\r\n    O
```

Figure 1: Example of a sentence extracted from the corpus.

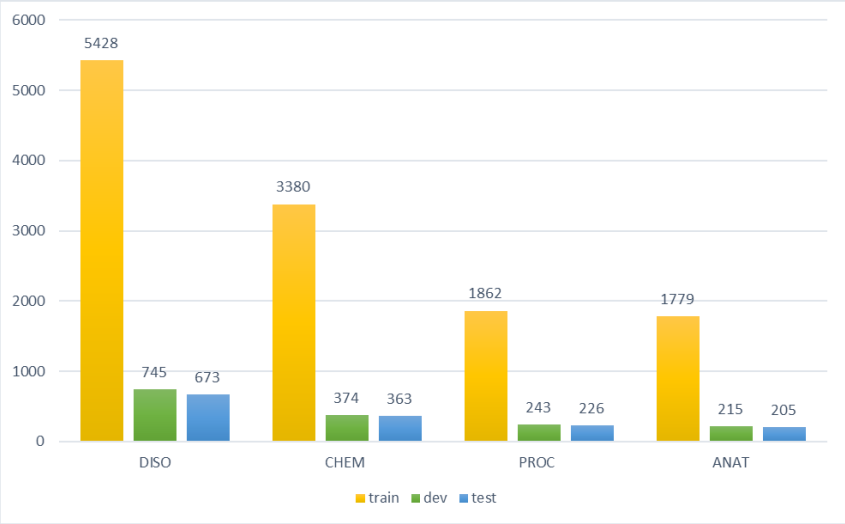

Figure 2: Label distribution in the train, dev and test data.

4 System architecture

For the purposes of this work we constructed a state-of-the-art system for NER in the Romanian biomedical domain using contextualized embeddings. Previous work relied solely on static embeddings. In order to compare the newly proposed system with previous approaches we also trained a system making use of static embeddings. This was necessary since existing systems were trained on the previous, smaller, version of the corpus, hence no direct comparison was possible. Comparison with older models is further made difficult by the introduction of new terms (such as COVID-related). The results for both systems, using static and contextual embeddings, are described in Section 5.

NER systems making use of transformer-based models usually obtain the numeric representations associated with input tokens which are then fed into a linear layer. Finally a classification head is used to obtain the predictions. In our approach, we employed an additional layer inspired by the biological process of lateral inhibition. In neurobiology, this process is defined as the capacity of an excited neuron to reduce the activity of its neighbors. This new layer is inserted after the embeddings calculation and before the linear layer.

To emulate the way inhibitory inter-neurons function, an embedding dimension value is either allowed to pass unchanged to the next layer or set to zero, depending on the other values. The forward pass calculation is given in Equation 1, where X is the layer's input vector, associated with a token embedding representation, Diag represents a matrix with the diagonal set to the vector given as parameter, ZeroDiag is the matrix with the value zero on the diagonal, and W and B represent the weights and bias. Θ is the Heaviside function, described in Equation 2.

$$F(X) = X * Diag(\Theta(X * ZeroDiag(W) + B)) \quad (1)$$

$$\Theta(x) = \begin{cases} 1, x > 0 \\ 0, x \leq 0 \end{cases} \quad (2)$$

The problem of computing a derivative for the Heaviside function in the backward pass was overcome by approximating the Heaviside function with the sigmoid function using a scaling parameter

as suggested by Wunderlich and Pehle (2021). This approximation was used only in the backward pass, while in the forward pass the Heaviside function was used as it is. This approximation technique is also known as surrogate gradient learning (Neftci et al., 2019) allowing the use of a non-differentiable function in the forward pass (e.g. the Heaviside function) while using a different function for approximating the derivative in the backward pass. The derivative of the sigmoid function is given in Equation 4, where $\sigma(x)$ is the same as in Equation 3.

$$\sigma(x) = \frac{1}{1 + e^{-kx}} \tag{3}$$

$$\sigma'(x) = k\sigma(x)\sigma(-x) \tag{4}$$

5 Results

Lee et al. (2020) has shown that contextual word representations trained on domain-specific biomedical corpora, such as BioBERT, largely outperforms BERT (Devlin et al., 2019) and previous state-of-the-art models in a variety of biomedical text mining tasks, including NER. However, for the Romanian language there is currently no contextual embedding model trained specifically on biomedical text. Therefore, for the purpose of this work we were forced to use either static word embeddings, trained on domain-specific data, or general-domain contextual models.

With regard to static word embedding models, Păiș and Tufiș (2018) have trained and published models using the Representative Corpus of Contemporary Romanian Language (CoRoLa) (Barbu Mititelu et al., 2019; Cristea et al., 2019). The authors have shown that due to the nature of the CoRoLa corpus, the models outperform existing ones, such as WikiPedia based models. Furthermore, the CoRoLa-based embeddings were previously used in constructing a Romanian language legal-domain NER system (Păiș et al., 2021; Păiș and Mitrofan, 2021b).

Following the approach of Păiș and Mitrofan (2021a), we wanted to explore the impact of using a combination of different word embeddings. Hence, we trained domain-specific static word representations on the BioRo corpus (Mitrofan and Tufiș, 2018), using the FastText toolkit[3] (Bojanowski et al., 2017). The resulting models can be downloaded from the RELATE platform[4] (Păiș et al., 2020).

Model	F1
CoRoLa	76.85
BioRo_5	**77.31**
BioRo_20	75.78
CoRoLa + BioRo_5	77.02

Table 2: Overall F1 scores using static word embedding models

We employed a recurrent neural network architecture, using Long Short Term Memory (LSTM) cells, representing tokens by means of pre-trained word embeddings with additional character embeddings, computed on the fly. The actual prediction is performed by a final Conditional Random Fields (CRF) layer. Implementation was realized using the NeuroNER[5] package (Dernoncourt et al., 2017).

The results obtained using the static word representation models are given in Table 2. The domain-specific word embeddings BioRo_5 achieves the best F1 score of 77.31%. This model contains representations for words appearing at least 5 times in the BioRo corpus. This result was expected since domain-specific models are known to perform better than general models. However, we were expecting to see an improvement when using the combination of general and domain-specific models. We assume the result given in Table 2 is due to the CoRoLa model being too general, while the SiMoNERo corpus contains only specialized text.

Contextual word representation models specifically created for Romanian language include Romanian BERT (Dumitrescu et al., 2020), RoBERT (Masala et al., 2020), JurBERT (Masala et al., 2021). Nevertheless, these models were not trained on biomedical text. However, Romanian language is also present in multilingual models, such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Lewis et al. (2020) recently showed that RoBERTa-based models produce state-of-the-art results in biomedical and clinical tasks. Therefore, we explored using the XLM-RoBERTa model with the system described in Section 4.

The results for the newly introduced system are

[3] https://fasttext.cc/

[4] https://relate.racai.ro/index.php?path=lrlt/models

[5] http://neuroner.com/

Entity	P	R	F1
ANAT	84.04	87.75	85.85
CHEM	82.64	89.25	85.82
DISO	84.72	86.35	85.53
PROC	76.47	77.69	77.08
Overall	**82.73**	**85.87**	**84.27**

Table 3: Results obtained with the proposed system

presented in Table 3. As expected, contextualized embeddings provide better results, even though they are not produced from domain-specific text. The hardest entity to predict is PROC, which we consider to be a result of the relatively low number of examples present in the corpus, given the complexity associated with this entity type (see Table 1). The ANAT entity type is the least common entity type, yet it is predicted to have the highest F1 score. We consider this to happen due to the reduced complexity of the entity type.

We further compared the results obtained with the newly introduced lateral inhibition layer with the same system without this layer. The overall F1 score was 83.42%, thus the new layer accounted for 0.85% improvement, under similar conditions (the same dataset split, the same contextual embeddings model, similar hyper-parameters).

6 Conclusion

This paper introduced a neural named entity recognition system adapted for the Romanian biomedical domain. It employed the new extended version of SiMoNERo corpus for training and evaluation. The proposed NER system uses pre-trained contextual embeddings, XLM-RoBERTa, and an inhibitory layer, inspired by the biological process of lateral inhibition. This work can make significant contribution in helping researchers interested in domain-specific NER both for Romanian and for other languages. In addition, the lateral inhibition mechanism has the potential to be applied in other tasks as well. Currently, it has been successfully applied in our system that participated in the SemEval 2022 shared task on Multilingual Complex Named Entity Recognition (MULTICONER)[6].

The resulting NER system is available for online usage through the RELATE platform[7]. The source code is freely available from our GitHub repository[8].

7 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 and part in the context of "Enrich4all" project, Action 2020-EU-IA-0088 funded by the European Union's Connecting Europe Facility 2014-2020 CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278547.

References

Sophia Ananiadou, Carol Friedman, and Jun'ichi Tsujii. 2004. Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6):393–395.

Verginica Barbu Mititelu, Dan Tufiș, Elena Irimia, Vasile Pais, Radu Ion, Nils Diewald, Maria Mitrofan, and Onofrei Mihaela. 2019. Little strokes fell great oaks. creating corola, the reference corpus of contemporary romanian.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tiberiu Boroș, Ștefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. Nlp-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179.

Ronald A Cohen. 2011. Lateral inhibition. *Encyclopedia of Clinical Neuropsychology, Springer, New York*, pages 1436–1437.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dan Cristea, Nils Diewald, Gabriela Haja, Cătălina Mărănduc, Verginica Barbu Mititelu, and Mihaela Onofrei. 2019. How to find a shining needle in the haystack. querying corola: solutions and perspectives. *Revue Roumaine de Linguistique*, No./Issue 3(3):279–292.

[6] https://multiconer.github.io/

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

[8] https://github.com/racai-ai/RNER

Ricardo C Cury, Istvan Megyeri, Tony Lindsey, Robson Macedo, Juan Batlle, Shwan Kim, Brian Baker, Robert Harris, and Reese H Clark. 2021. Natural language processing and machine learning for detection of respiratory illness by chest ct imaging and tracking of covid-19 pandemic in the united states. *Radiology: Cardiothoracic Imaging*, 3(1):e200596.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Ştefan 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*, pages 4324–4328.

Radu Ion. 2007. *Word sense disambiguation methods applied to English and Romanian*. Ph.D. thesis, PhD thesis (in Romanian). Romanian Academy, Bucharest.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mihai Masala, Stefan 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.

Verginica Barbu Mititelu and Maria Mitrofan. 2020. The romanian medical treebank-simonero. *ISSN 1843-911X*, page 7.

Maria Mitrofan. 2017. Bootstrapping a romanian corpus for medical named entity recognition. In *RANLP*, pages 501–509.

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

Maria Mitrofan, Verginica Barbu Mititelu, and Grigorina Mitrofan. 2019. Monero: a biomedical gold standard corpus for the romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 71–79.

Maria Mitrofan and Dan Tufiș. 2018. Bioro: The biomedical corpus for the romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1192–1196.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. 2019. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

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

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

Vasile Păiș and Maria Mitrofan. 2021a. 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 Maria Mitrofan. 2021b. Towards a named entity recognition system in the romanian legal domain using a linked open data corpus. In *Workshop on Deep Learning and Neural Approaches for Linguistic Data*, pages 16–17, Skopje, North Macedonia.

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.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.

Leon Weber, Jannes Münchmeyer, Tim Rocktäschel, Maryam Habibi, and Ulf Leser. 2020. Huner: improving biomedical ner with pretraining. *Bioinformatics*, 36(1):295–302.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Timo C. Wunderlich and Christian Pehle. 2021. Event-based backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1):12829.

Ying Xiong, Yuanhang Huang, Qingcai Chen, Xiaolong Wang, Yuan Nic, and Buzhou Tang. 2020. A joint model for medical named entity recognition and normalization. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.