Legal Named Entity Recognition with Multi-Task Domain Adaptation

Răzvan-Alexandru Smădu[1], **Ion-Robert Dinică**[1], **Andrei-Marius Avram**[1],
Dumitru-Clementin Cercel[1], **Florin Pop**[1,2], **Mihaela-Claudia Cercel**[3]
[1]University Politehnica of Bucharest, Faculty of Automatic Control and Computers
[2]National Institute for Research & Development in Informatics - ICI Bucharest, Romania
[3]First District Court of Buftea
{razvan.smadu,ion_robert.dinica}@stud.acs.upb.ro
{andrei_marius.avram}@stud.acs.upb.ro
{dumitru.cercel,florin.pop}@upb.ro

Abstract

Named Entity Recognition (NER) is a well-explored area from Information Retrieval and Natural Language Processing with an extensive research community. Despite that, few languages, such as English and German, are well-resourced, whereas many other languages, such as Romanian, have scarce resources, especially in domain-specific applications. In this work, we address the NER problem in the legal domain from both Romanian and German languages and evaluate the performance of our proposed method based on domain adaptation. We employ multi-task learning to jointly train a neural network on two legal and general domains and perform adaptation among them. The results show that domain adaptation increase performances by a small amount, under 1%, while considerable improvements are in the recall metric.

1 Introduction

Legal is one of the domains where NER plays a central role, especially in document processing, where it is used for identifying key elements like the court name, the name of the parties in a case, or the case number (Skylaki et al., 2020). In recent years, interest has grown in the research community for performing various tasks on legal documents, known as LegalAI (Zhong et al., 2020).

One direct application of these extracted named entities is document organization and search. However, they can be further incorporated into other systems like document anonymization, judgement prediction, or case summarization, offering additional insights to legal professionals (Zhong et al., 2020; Bansal et al., 2019).

Although still considered under-resourced, Romanian is one of the languages that has seen a recent expansion with the introduction of two Bidirectional Encoder Representation from Transformers (BERTs) (Devlin et al., 2019) trained on Romanian text (Dumitrescu et al., 2020; Masala et al., 2020), three named entity corpora (Dumitrescu and Avram, 2020; Păiș et al., 2021b; Mitrofan and Tufiş, 2018), over three hundred hours of publicly-available transcribed speech (Georgescu et al., 2020; Wang et al., 2021), and a benchmark that tracks the progress of various Romanian NLP tasks (Dumitrescu et al., 2021). In addition, domain adaptation research showed that we could perform knowledge transfer between datasets using effective methods in both supervised (Yue et al., 2021) and unsupervised (Ganin and Lempitsky, 2015) settings.

In this work, we want to take advantage of these recent developments and explore the area of domain adaptation with a task discriminator on the Romanian language. On a more granular level, we experiment with domain adaptation from the general to the legal domain, using the Romanian Named Entity Corpus (RONEC) (Dumitrescu and Avram, 2020) as a reference and Romanian Legal NER corpus (LegalNERo) (Păiș et al., 2021b) as a target.

Our proposed neural architecture employs multiple components. A pre-trained BERT layer generates the feature representation. Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is utilized in bidirectional configuration to capture both left-to-right and right-to-left dependencies. Conditional Random Fields (CRFs) (Lafferty et al., 2001) generate the predictions based on the conditional probability of the sequence. We employ multi-task learning (Changpinyo et al., 2018) to train the model on legal and general domains jointly, and on top of that, we apply domain adaptation as described by Ganin and Lempitsky (2015), but in a supervised setting.

Furthermore, to explore the robustness of our approach in another but better-resourced language, we apply the same methodology to German and investigate domain adaptation from GermEval 2014 (Benikova et al., 2014) to German Legal NER

December 8, 2022 ©2022 Association for Computational Linguistics

(German LER) (Leitner et al., 2020). Ultimately, we evaluate our approach through visualizations and analysis of the predictions.

We summarize the contributions as follows:

- We propose a multi-task domain adversarial model that is jointly trained on two domains, namely general and legal;
- We evaluate the performances of our approach, the quality of the predictions, and propose a way of visualizing the embedding space of the named entities;
- To the best of our knowledge, we are the first to experiment with domain adaptation in the legal NER.

2 Related Work

Named Entity Recognition. In the supervised setting, Lample et al. (2016) introduced LSTM-based neural architectures that do not rely on domain-specific resources or hand-crafted features. Both character-level and word-level representations passed through LSTM and CRF layers proved effective in NER.

Since Transformers (Vaswani et al., 2017) became popular in NLP, BERT-based approaches were evaluated on NER tasks, proving the effectiveness of the contextualized word embeddings and transformer word representations (Souza et al., 2019; Dai et al., 2019; Jiang et al., 2019; Liu et al., 2020; Syed and Chung, 2021). These methods combine the previous neural components (i.e., LSTM, BERT, and CRF) with deep learning techniques such as transfer learning (Weiss et al., 2016), active learning (Cohn et al., 1996), and domain adaptation. Pointer Generator Networks (See et al., 2017) were also employed in NER by Skylaki et al. (2020), showing that the proposed method achieves better results when compared to BERT-based and LSTM-based models.

Often, NER is jointly addressed along with relation extraction (Feldman and Rosenfeld, 2006; Nasar et al., 2021). While NER is usually solved using recurrent neural networks, relation extraction can be handled using convolutional (Zheng et al., 2017) and feed-forward layers (Bekoulis et al., 2018; Bhatia et al., 2019; Shi and Lin, 2019).

To generate labels for new types of entities and relation extraction in automated systems, distant supervision (Mintz et al., 2009) is utilized based on a dataset of entities. Improvements rely on introducing a reinforcement learning module in the tagger that selects clean data for the model architecture during training (Yang et al., 2018).

NER tasks can become challenging when entities are nested; often, they address flat NER. Generally, classical approaches do not consider nesting and treat this task as dependency parsing (Yu et al., 2020). The method relies on embeddings generated via BERT for word-based embeddings and convolutional layers for char-based embeddings. Feed-forward layers and a biaffine model (Dozat and Manning, 2017) predict the entity spans.

Legal NER. Previously, classical machine learning techniques such as Support Vector Machines, Naive Bayes, and ontologies were utilized to detect named entities from legal documents (Dozier et al., 2010; Bruckschen et al., 2010; Cardellino et al., 2017; Glaser et al., 2018). New datasets started to emerge in the legal domain since legal is one of the domains that received little attention (Leitner et al., 2019). Methods based on domain-specific embeddings and LSTMs combined with CRFs were utilized in multiple languages, such as English (Chalkidis et al., 2019), German (Leitner et al., 2020), Romanian (Păiș et al., 2021a), and Portuguese (Luz de Araujo et al., 2018). Barriere and Fouret (2019) employed a two-learning-step approach that first trains a model on the NER task, which then creates features for training a second neural network model. This approach was evaluated on French legal documents, showing a significant reduction in the F1-score error.

Domain Adaptation. The domain adaptation setting aims to reduce the domain gap between the source and target data distributions. This technique takes advantage of the knowledge of well-resourced domains and transfers it to downstream tasks with fewer resources. Jia and Zhang (2020) approached the cross-domain NER via multi-task learning and a variation of the LSTM cell. Their approach evaluated on few-shot datasets showed significant improvements over other multi-task learning methods. In the cross-domain setting, Liu et al. (2021) tested multiple model architectures based on BERT, LSTM, and CRFs. Their experiments suggested that domain-adaptive pre-training can enhance both span-level and token-level performances.

Various transfer learning and fine-tuning techniques, such as parameter initialization and multi-

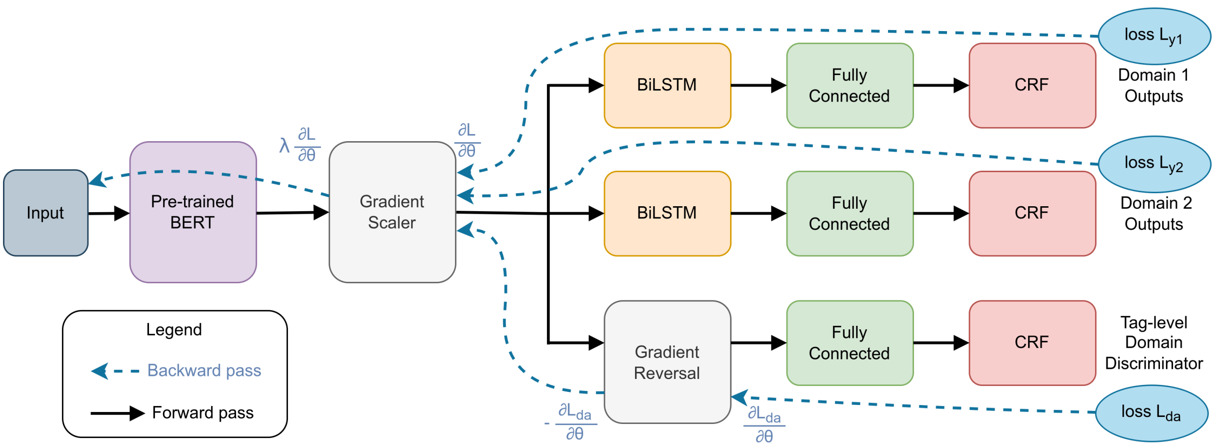


Figure 1: The proposed model architecture.

task learning, can be employed to reduce the training overhead of neural architectures (Lin and Lu, 2018). Bekoulis et al. (2018) enhanced the LSTM-CRF architecture by adding adversarial training through adversarial perturbation in the embedding space. Virtual Adversarial Training (VAT) (Miyato et al., 2019) was combined with the LSTM-CRF model in both supervised and semi-supervised settings. In the minimization objective, they introduced the Kullback-Leibler divergence computed between estimated labels of original and adversarial examples (Chen et al., 2020). VAT significantly improved performance over baseline models.

A language discriminator is added in the multi-lingual setting to perform adversarial learning (Chen et al., 2021). The goal of the discriminator is to force the feature encoder to learn language-invariant features. Moreover, domain adaptation was achieved using latent semantic association such that the same concepts from different domains should be semantically similar (Guo et al., 2009).

3 Approach

3.1 Neural Network Architecture

Inspired by the previously mentioned works, we based our neural network architecture on a domain adaptation technique via multi-task learning. We consider a two-domain model, which jointly trains, in a supervised fashion, on two datasets from domains characterized by a domain shift. Figure 1 presents the complete model architecture.

Each domain is associated with a branch in the model architecture while sharing the feature encoder. We utilize contextualized BERT embeddings to generate the feature space. Two branches and a domain discriminator process the BERT's output. The transformer model is pre-trained on the language of the datasets we use and follows a fine-tuning approach during training, such that we do not change too much the embedding space, but it is still subject to domain adaptation.

Implementation-wise, we introduce a gradient scaler layer that scales down the gradients during back-propagation by a factor γ, similar to a learning rate. We apply a scheduler that increases this learning rate over time:

$$\gamma^* = \frac{1}{1 + e^{\gamma(-2p+1)}} \tag{1}$$

where γ^* is the learning rate at the current progress rate $p \in [0, 1]$, and e is Euler's number. Our intuition is that at the beginning of the training, we want to avoid affecting the pre-trained Transformer model's weights since the higher-level layers are not trained, and we enable fine-tuning after some training steps.

Each domain branch comprises a BiLSTM (Bidirectional LSTM), followed by a fully connected layer and a CRF output layer. We use BiLSTMs since these are more resilient to gradient vanishing (Hochreiter and Schmidhuber, 1997) while capturing feature dependencies from both left-to-right and right-to-left directions, and CRFs to model the conditional probability distribution of the input sequence. Lastly, we introduce a discriminative branch linked to the shared embedding encoder via a gradient reversal layer (Ganin and Lempitsky, 2015), having a linear layer followed by CRF. The motivation for the usage of domain adaptation is presented in Section 3.3.

3.2 Conditional Random Field

CRFs (Lafferty et al., 2001) are discriminative models based on undirected graphs, modelling the conditioned probability of labels obeying the Markov property relative to the dependency graph, given the input (i.e., $P(y|X)$) (Sutton and McCallum, 2012). The input of the CRF is a sequence of features of the input sequence $X = (x_1, x_2, ..., x_n)$, being output by the last fully connected layer. The output sequence of the CRF is a label y from the set of all possible classes K. For each pair of input sequence labels, its score is defined as:

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{2}$$

where A is the transition score matrix of size $K \times K$, and P are the output scores generated by the last fully connected layer of size $n \times K$. Probability of the sequence y is defined as a softmax over all scores:

$$p(y|X) = \frac{\exp\left(s(X, y)\right)}{\sum_{y' \in Y_X} \exp\left(s(X, y')\right)} \tag{3}$$

with Y_X being the set of all possible tag sequences for a sequence X. Compared with the softmax activation function, the CRF can handle sequential dependencies.

To determine the predictions of the input sequence, we run the Viterbi algorithm (Forney, 1973), which extracts the tags y^* with the maximum score:

$$y^* = \arg\max_{y' \in Y_X} s(X, y') \tag{4}$$

The optimization process is based on maximizing the likelihood:

$$\log p(y|X) = s(X, y) - log(\sum_{y' \in Y_X} \exp\left(s(X, y')\right)) \tag{5}$$

It allows us to combine CRFs with neural network models, where the loss function is the negative log-likelihood (i.e., $L_{CRF} = -\log p(y|X)$), which is optimized using gradient-based methods.

3.3 Optimization in Domain Adaptation Setting

One of the most influential works is in the unsupervised domain adaptation setting (Ganin and Lempitsky, 2015), which aims to reduce the domain shift by introducing a domain discriminator. Similar to how Generative Adversarial Networks (Goodfellow et al., 2014) work, the domain discriminator learns indistinguishable feature representations between different domains. Therefore, it minimizes the loss function concerning the labels while maximizing the error rate of the domain discriminator. This minimax game is formalized as follows:

$$\hat{\theta}_f, \hat{\theta}_y = \arg\min_{\theta_f, \theta_y} L(\theta_f, \theta_y, \hat{\theta}_d) \tag{6}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} L(\hat{\theta}_f, \hat{\theta}_y, \theta_d) \tag{7}$$

where L is the loss function, θ_f are the parameters of the feature generator, θ_y are the parameters of the label predictor, and θ_d are the parameters of the domain discriminator. The variables with a hat are fixed during optimization. The empirical loss function is the difference between the prediction loss L_y and domain adaptation loss L_d:

$$L = L_y - \lambda L_{da} \tag{8}$$

where λ is a hyperparameter that controls the level of domain adaptation during training. This optimization problem is implemented by linking the discriminator to the feature extractor via a gradient reversal layer that negates the gradient during back-propagation while solving:

$$L = L_y + \lambda L_{da} \tag{9}$$

In this paper, we utilize the domain adaptation at the tag level, meaning that the discriminator learns specific feature representations based on the context of each tag. Therefore, we use a CRF layer to model the sequence of constant values among the same domain (for example, a sequence of 1s for the first domain and 2s for the second domain). This also motivates why we use a gradient scaler after the BERT layer.

When performing the feed-forward step, a batch containing examples from both domains is passed through the model. First, we utilize the samples from the first domain and accumulate gradients computed for the loss associated with the first domain L_{y_1} and domain discriminator L_{da_1}. Then, we pass the examples from the second domain and accumulate the gradients for the second domain output L_{y_2} and the domain discriminator L_{da_2}. Next, we perform gradient updates and repeat the training procedure for all batches in the training set. See Figure 1 for how the gradients are propagated throughout the network.

We minimize the negative log-likelihood loss for each branch, computed as described in Section 3.2. The total loss is formalized below:

$$L_{total} = L_{y_1} + L_{y_2} - \lambda(L_{da_1} + L_{da_2}) \quad (10)$$

We vary λ according to the same Equation (1) but scaled by a constant α. Hence, we enable more domain adaptation over time at a progress rate p:

$$\lambda_p = \alpha\left(\frac{2}{1 + e^{-\beta p}} - 1\right) \quad (11)$$

where α defines the upper boundary of the function, and β controls the widening of the sigmoid function.

During training, we observed that the adversarial loss starts to increase after a period of training. At the same time, the discriminator performs poorly (that means the discriminator becomes unable to distinguish among features). Subsequently, this also hinders the performance on the other tasks, and limiting domain adaptation proved to yield better results. Another observation we made is that by negating the loss term of the domain discriminator, instead of using $+L_{da}$ (further referenced as ADAL), we utilize $-L_{da}$ during optimization (further referenced as SDAL). The performances considerably improved when compared with classical domain adaptation.

4 Experiments

4.1 Datasets

We evaluate our approach on datasets from general and legal domains, Romanian and German, respectively. We utilize the splits provided; where these were not provided, we randomly split the datasets into train/test/validation, using 80%-20%-20% ratios.

RONEC (Named Entity Corpus for the Romanian language) (Dumitrescu and Avram, 2020)[1] is an open-source dataset, currently at version 2.0, containing 0.5M tokens within 12,330 annotated sentences extracted from newspapers. The total number of entities annotated in the RONEC v2.0 corpus is 80,283 from 15 distinct classes, inspired by the OntoNotes5 (Weischedeld et al., 2013) and ACE (Doddington et al., 2004) datasets. The dataset is available under CoNLL-U format[2], using the BIO annotation schema (Lample et al., 2016). The second version was annotated by `termene.ro`[3]. The dataset is split into 9,000 sentences for training, 1,330 sentences for validation, and 2,000 for testing. The entity classes are roughly evenly balanced among the splits. This dataset also has version 1.0 available but was not utilized during experiments.

LegalNERo (Păiș et al., 2021a)[4] is a named entity corpus proposed for the Romanian language by researchers from the Romanian Institute of Artificial Intelligence. This dataset was annotated by five human annotators and consists of 370 documents extracted from the MARCELL-RO corpus (Tufiș et al., 2020). The dataset contains 8,284 sentences and a total of 13,614 entities. The entity classes considered in this dataset are the following: Person, Location, Organization, Time, and Legal Ref. The dataset is available in the CoNLL-U Plus format, annotated using the BIO schema.

GermEval 2014 (Benikova et al., 2014)[5] is a dataset proposed at the KONVENS workshop that introduces an extended set of tags compared with previous works. In summary, it contains 31,300 annotated sentences, consisting of a total of 590,000 tokens and 41,124 entities from four main classes (person, location, organization, and other), as well as derivations and parts of named entities for each of the main classes (there are 12 classes in total). The dataset is divided into three sets: 24,000 sentences for training, 2,200 sentences for validation, and 5,100 sentences for testing, all being provided in the CoNLL-U format, following the BIO schema.

LER (Legal Entity Recognition) (Leitner et al., 2019)[6] contains 750 court decisions from Germany, which were published on an online portal (i.e., "Rechtsprechung im Internet"; in eng. "Jurisprudence on the Internet"). The dataset has 66,723 sentences, which consists of 53,632 annotated named entities. This dataset is available in the CoNNL-U format, using the BIO annotation schema. The dataset consists of seven categories for the named entities (i.e., person, location, organization, legal norm, case-by-case regulation, court decision, and legal literature), divided into 19 fine-

[1] `https://github.com/dumitrescustefan/ronec`

[2] `https://universaldependencies.org/format.html`

[3] `https://termene.ro`

[4] `https://zenodo.org/record/4772095`

[5] `https://sites.google.com/site/germeval2014ner/`

[6] `https://github.com/elenanereiss/Legal-Entity-Recognition`

grained classes.

4.2 Data Preprocessing

Extracting named entities from documents can be cast to a tagging problem. Each word follows the BIO schema (Lample et al., 2016) (i.e., the beginning of entities are labelled with B, inside tokens are labelled with I, and outside of entities are annotated with O). The labels are numerically encoded such that each label indicate the BIO tag and its class.

In this work, we employ Transformer representations, and we utilize the pre-trained BERT tokenizer (Sennrich et al., 2016) on the language we train the model to generate the input tokens. Since the goal is to keep a small enough vocabulary, some words are split among multiple tokens. In this case, we consider a `NULL` tag that indicates the token is an inside subword (for example, when using BERT tokenizer, these tokens start with `'##'`). Each sample consists of a sentence. If the sentence length after tokenization is longer than the maximum sequence length for BERT, then we split the sentence into multiple examples.

4.3 BERT Embedding Representation

We use the language-specific pre-trained BERT model since we deal with multiple languages. For the Romanian language, we use the pre-trained Romanian BERT model (Dumitrescu et al., 2020), which was trained on three corpora, namely OPUS (Tiedemann, 2012), OSCAR (Suárez et al., 2019), and Romanian Wikipedia, in total representing 15.2GB of processed data. We employed the cased base model. For the German language, we utilized the German BERT model (Chan et al., 2020), which was pre-trained on four datasets (i.e., the German version of OPUS, OSCAR, and Wikipedia, at which it is added the Open Legal Data (Ostendorff et al., 2020) – a dataset for legal domain), worth over 150GB of data. We used the cased base model in our experiments since some words (such as nouns) are spelt with capital letters at the beginning of the words.

4.4 Baselines

We compare our approach with simplified versions of the proposed architecture, considered baseline. They consist of a BERT transformer, a BiLSTM layer, a fully connected layer, and a CRF layer for generating the probability distribution for the tokens. We trained this architecture on all four datasets and followed the same training procedure as the proposed method. During training, we set the BERT model to be fine-tuned to improve the embedding generation on the downstream task.

4.5 Experimental Setup

We trained all models on TPUv3-8[7] provided by Kaggle[8] for free. We used a batch size between 4 and 16 per TPU, and trained the baseline models for at most 10 epochs. The learning rate was varied using a linearly decreasing scheduler, with the warm-up proportion set to 1%. The maximum learning rate was set to 0.002, and the minimum value was attained at the last epoch. In all cases, we used a gradient scaler value of 1e-5. To reduce overfitting, we utilized the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 1e-5. In addition, we employ gradient clipping of magnitudes greater than 2.0. For tokenizer, we set the maximum sequence length to 200. For domain adaptation setup, similar hyperparameters were employed. In addition, we set the maximum epoch to be 20 while keeping the best-performing checkpoint for evaluation, and the domain adaptation hyperparameter α was set to 0.1. In contrast, β was set to 10.

4.6 Evaluation Metrics

We assess the performance of the models in terms of negative log-likelihood computed as described in Section 3.2, and F1-score at the entity level (Yadav and Bethard, 2018; Dumitrescu et al., 2020) from four metrics: Entity Type, Partial, Exact, and Strict, computed as follows[9]:

$$P = \frac{Correct}{Correct + Incorrect + Partial + Spurius} \tag{12}$$

$$R = \frac{Correct}{Correct + Incorrect + Partial + Missing} \tag{13}$$

$$F1 = \frac{2PR}{P + R} \tag{14}$$

where $Correct$ represents correctly predicted entities; $Incorrect$ are the incorrectly predicted labels by the system; $Partial$ are the partially correct detected annotations; $Missing$ are the golden labels not detected by the model; $Spurius$ are the entities detected by the model, but they are not in the gold set.

[7] https://cloud.google.com/tpu

[8] https://www.kaggle.com/

[9] https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/

5 Results

This section presents results on baseline models and domain adaptation. We present the precision, recall, and F1-scores at the entity level (strict measures). In the end, we provide t-SNE (van der Maaten and Hinton, 2008) visualizations of the embedding space on the feature space along with the limitations of the current approach. In Appendix A, we present more detailed results.

5.1 Baselines

For the baseline models, we present the results in Table 1. Results are obtained in the following configurations: RONEC - LegalNERo, and GermEval 2014 - LER. We present the negative log-likelihood averaged per token (NLL) as absolute values (the lower, the better) and F1-scores as percentages (the higher, the better). The German LER model achieves the highest F1-scores and the lowest NLL absolute value (0.0676) since this is the largest dataset we used. On the other side of the spectrum, the model trained on LegalNERo achieves the smallest F1-scores, with only 80.3%, being at the same time the smallest dataset. On RONEC, we observed higher scores for Entity Type and F1-Partial, meaning that the models partially identified entities in the text, while the exact boundaries and, in some cases, the types of the entities were misidentified. However, on this dataset, the model achieves the highest NLL score. On the GermEval 2014 dataset, the model reaches the best score when identifying partial entities. In contrast, the smallest score is achieved when determining the exact match of the boundaries and entity types.

Dataset	NLL	Ent. Type	F1-Partial	F1-Exact	F1-Strict
RONEC	0.1121	90.51	91.22	89.52	87.51
LegalNERo	0.0900	86.69	85.21	81.07	80.30
GermEval 2014	0.0684	84.60	88.40	84.22	82.62
LER	0.0676	96.18	93.76	90.37	89.86

Table 1: Baseline results obtained on all datasets.

All models achieve over 85% in F1-score in partially identifying entities while ignoring their type. At the same time, we observe that the performances drop by almost 5% when the model has to identify the exact boundaries and entity type. In general, we observe the following patterns inspecting the outputs of the baseline models:

- In the case of nested entities, the models predict the inside entities but not the whole one;
- The precision is lower compared with recall, meaning that the models predict non-existing entities, while those that correspond to ground truth are correctly identified and classified;
- Some classes are repeatedly misclassified (for example, on LegalNERo, organizations are predicted as persons, or in RONEC, events are predicted as organizations);
- Invalid boundaries, as observed earlier by the drop of 5% in strict metric compared with the partial metric.

5.2 Domain Adaptation on Different Domains

Applying domain adaptation, we experimented with two configurations: first, in which we add the domain adaptation loss $+L_{da}$, and second, in which we subtract the domain adaptation loss $-L_{da}$.

Table 2 presents the results of the ADAL scenario. We observe the performances are similar to the baseline models, meaning that even if we perform domain adaptation, the input space's new latent structure does not help improve performances. In general, the results are slightly lower, by at most 3% in F1-score. In the case of LegalNERo, we see an improvement of 1% in strict and exact metrics. In almost all cases, the evaluation NLL score is larger than the baseline values, the exception being LER.

Dataset	NLL	Ent. Type	F1-Partial	F1-Exact	F1-Strict
RONEC	0.1561	88.14	89.49	87.45	84.51
LegalNERo	0.1239	86.58	85.79	82.05	81.24
GermEval 2014	0.0739	83.77	88.21	87.13	81.96
LER	0.0245	94.45	93.54	90.62	89.15

Table 2: Domain adaptation trained using ADAL.

On SDAL (see Table 3), we notice improvements of up to 3% along almost all datasets, except for LegalNERo, where the performance drops by 5%. We note that the highest score obtained on LER is 92.2%, GermEval 2014 is 85.57%, and RONEC is 87.10%. Compared with ADAL, these scores are higher by up to 4% in the case of the generic datasets.

Dataset	NLL	Ent. Type	F1-Partial	F1-Exact	F1-Strict
RONEC	0.1084	90.17	90.86	89.13	87.10
LegalNERo	0.0910	81.21	79.95	76.39	75.68
GermEval 2014	0.0666	86.91	90.66	89.87	85.57
LER	0.0168	96.27	95.27	93.07	92.30

Table 3: Domain adaptation trained using SDAL.

5.3 In-Dataset Domain Adaptation

We analyze the effects of applying domain adaptation inside the same dataset. In other words, we utilize the same train set for both domains while keeping different domain tags. The intuition is to enforce a feature representation that is more robust against small variations in the latent space due to the random initialization of both branches. Table 4 shows the results on all four datasets.

Dataset	NLL	Ent. Type	F1-Partial	F1-Exact	F1-Strict
RONEC	**0.2558**	88.39	90.07	88.42	85.59
RONEC	0.2620	**89.65**	**90.62**	**88.99**	**86.82**
LegalNERo	0.1498	84.86	77.67	66.05	63.46
LegalNERo	**0.1435**	**89.15**	**88.86**	**85.93**	**85.17**
GermEval 2014	0.1073	**85.31**	89.76	88.86	83.79
GermEval 2014	**0.1090**	85.08	**90.23**	**89.49**	**83.94**
LER	0.0208	95.59	93.96	91.05	90.40
LER	**0.0203**	**95.70**	**93.99**	**91.16**	**90.59**

Table 4: Domain adaptation on the same datasets: RONEC - RONEC, LegalNERo - LegalNERo, GermEval 2014 - GermEval 2014, and LER - LER.

We observe that almost consistently, one of the two task heads performs better than the other. In the case of LegalNERo, there is a considerable difference in performances between the two heads while keeping similar NNL scores. This indicates that small changes in per-tag measurements may have larger impacts on sequential measurements. Moreover, these results improve upon the baseline results, by a small margin, under 1%, on all datasets except RONEC, on which we observe performance degradation by at most 1%. In addition, we observe higher NLL scores, except on the LER dataset.

5.4 Effects of Domain Adaptation on the Feature Space

We generate t-SNE representations in the latent space of the model, outputted by the Transformer layer. The visualizations are generated at perplexity set to 30. We compare these representations between datasets and assess how well the model adapted to the changes in the data distributions.

Figure 2 shows the scenario on the Romanian datasets. The pre-trained BERT outputs are generated using the non-fine-tuned version of the pre-trained BERT model in the Romanian language. We observe that the pre-trained BERT outputs on RONEC are sparse and tend to form two blobs (i.e., a large one on the left and a smaller one on the right). On the LegalNERo dataset, the data points tend to cluster and not follow the same distribution as RONEC.

In the case of the German datasets (see Figure 3), we observe similar behaviors as on the Romanian datasets. The pre-trained BERT model on the German language outputs entities generates a latent space in which examples from the GermEval 2014 dataset are close to LER while tending to cluster into groups of points from the same dataset. This phenomenon is emphasized when domain adaptation is employed. In both ADAL and SDAL, there is a separation between datasets.

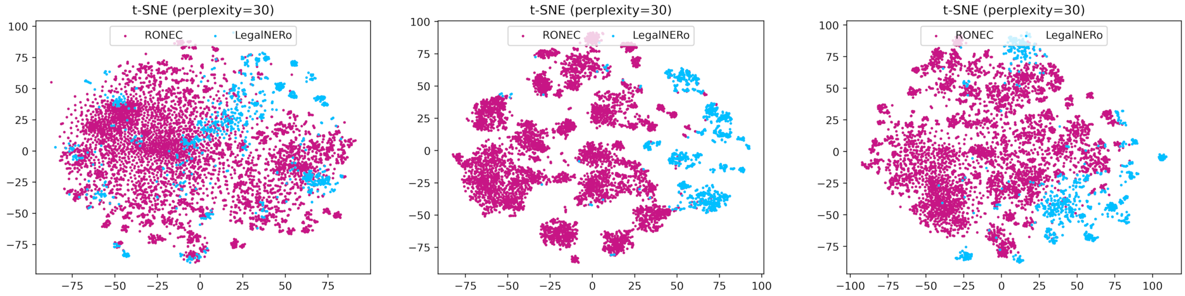
When analyzing ADAL, we can see that the data tend to form clusters and separation between examples from the two datasets. In the SDAL training scenario, we observe at a smaller degree the tendency of clustering. We can see that both datasets are separated, but the data points are not cluttering into some spots. Considering the performance differences, we may hypothesize that the feature predictor prefers sparser representations to compact ones.

Having linearly separable classes is the desired objective since this representation is much easier to be classified by the simpler classifiers. We cannot assume this is the case (i.e., linear separation) due to how t-SNE works. From these visualizations, we can deduce that combining both gradient reversal and gradient scaler layers are an effective way of shaping the latent space, if set appropriately.

Therefore, from this empirical observation, the sign of the domain adaptation loss term does influence the latent representation, more specifically, when considering the sparsity of the data. When we add the loss term and perform the optimization step in domain adaptation, we aim to minimize that term so that the domain classifier is trained like a regular neural network. At the same time, we maximize the loss function in the latent representation to generate similar feature distributions. In our experiments, we see the opposite in the multitask learning setup. The feature representation at the entity level tends to become separable and cluttered together. It is desired in the context of a label classifier since we want different features that are easily predictable for each class.

5.5 Limitations

Our proposed method has some limitations, especially during training. As previously mentioned, a more significant domain adaptation on the BERT architecture yields poor performances. This motivated us to introduce a gradient scaler layer and

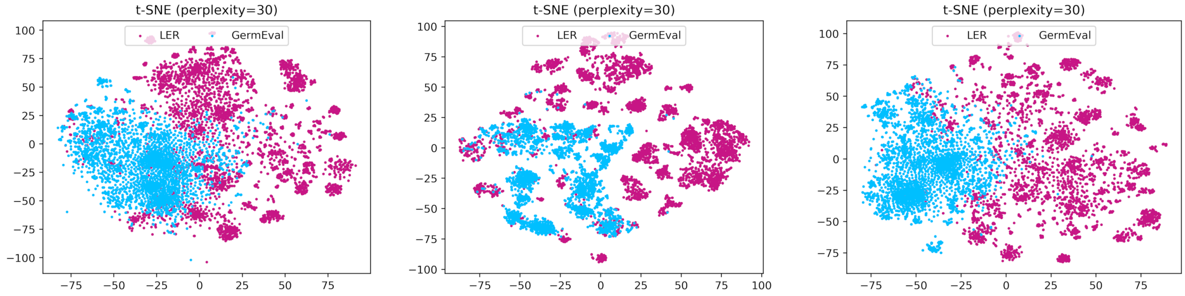


(a) Pre-trained BERT outputs

(b) ADAL training

(c) SDAL training

Figure 2: t-SNE visualizations of the embedding space on Romanian datasets for the baseline, ADAL, and SDAL.



(a) Pre-trained BERT outputs

(b) ADAL training

(c) SDAL training

Figure 3: t-SNE visualizations of the embedding space on German datasets for the baseline, ADAL, and SDAL.

the hyperparameter schedulers, thus reducing this effect. On the other hand, we analyzed the models' predictions and observed that some boundaries are incorrectly determined, thus affecting the scores. For example, in the Romanian language, the entity "Sanctităţii Sale Papa Francisc" (eng., "His Holiness Pope Francisc") is split into two entities (see Appendix A.5). Our method does not detect some entities on all datasets in a different context. Also, the model does not capture this variety in the dataset, which overfits this scenario. More discussions on limitations can be seen in the case study from Appendix A.4.

6 Conclusions and Future Work

We proposed a method based on multi-task domain adaptation in a cross-domain setting. The model architecture is based on contextualized word embeddings generated using BERT, LSTM, fully connected, and CRF layers. We evaluated our approach on two languages (i.e., Romanian and German) from two domains (i.e., general and legal). We observed minimal improvements in the German dataset while reducing performance on the Romanian legal dataset. More research should be conducted in this direction.

For future work, we strive to investigate the performance degradation further and analyze the effects of domain adaptation on the embedding space via t-SNE visualizations. In addition, we want to evaluate a cross-lingual setting, considering the cross-language BERT pre-trained model and performing domain adaptation between the same domain but different languages.

References

Neha Bansal, Arun Sharma, and R. K. Singh. 2019. A review on the application of deep learning in legal domain. In *Artificial Intelligence Applications and Innovations*, pages 374–381, Cham. Springer International Publishing.

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. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 327–332, Turku, Finland. Linköping University Electronic Press.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. Germeval 2014 named entity recognition shared task: companion paper. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*.

Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. 2019. Comprehend medical: a named entity recognition and relationship extraction web service. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851. IEEE.

Mırian Bruckschen, Caio Northfleet, DM Silva, Paulo Bridi, Roger Granada, Renata Vieira, Prasad Rao, and Tomas Sander. 2010. Named entity recognition in the legal domain for ontology population. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts*.

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.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Neural contract element extraction revisited. In *Workshop on Document Intelligence at NeurIPS 2019*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2965–2977, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020. Seqvat: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8801–8811.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 743–753, Online. Association for Computational Linguistics.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, pages 1–5. IEEE.

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

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. Named entity recognition and resolution in legal text. In *Semantic Processing of Legal Texts*, pages 27–43. Springer.

Stefan 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.

Stefan 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.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Ronen Feldman and Benjamin Rosenfeld. 2006. Boosting unsupervised relation extraction by using NER. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 473–481, Sydney, Australia. Association for Computational Linguistics.

G.D. Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, and Corneliu Burileanu. 2020. Rsc: A romanian read speech corpus for automatic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6606–6612.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. 2009. Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 281–289.

Christian Hänig, Stefan Thomas, and Stefan Bordag. 2014. Modular classifier ensemble architecture for named entity recognition on low resource systems.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional lstm for ner domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5906–5917.

Shaohua Jiang, Shan Zhao, Kai Hou, Yang Liu, Li Zhang, et al. 2019. A bert-bilstm-crf model for chinese electronic medical records named entity recognition. In *2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pages 166–169. IEEE.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany. Springer. 10/11 September 2019.

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.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022, Brussels, Belgium. Association for Computational Linguistics.

Mingyi Liu, Zhiying Tu, Zhongjie Wang, and Xiaofei Xu. 2020. Ltp: a new active learning strategy for bert-crf based named entity recognition. *arXiv preprint arXiv:2001.02524*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *AAAI*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer.

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.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

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

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1).

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 385–388.

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

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

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Stavroula Skylaki, Ali Oskooei, Omar Bari, Nadja Herger, and Zac Kriegman. 2020. Named entity recognition in the legal domain using a pointer generator network. *CoRR*, abs/2012.09936.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373.

Muzamil Hussain Syed and Sun-Tae Chung. 2021. Menuner: Domain-adapted bert based ner approach for a domain with limited dataset and its application to food menu domain. *Applied Sciences*, 11(13).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. 2020. Collection and annotation of the Romanian legal corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2773–2777, Marseille, France. European Language Resources Association.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Ralph Weischedeld, Martha Palmerd, Mitchell Marcusd, Eduard Hovyd, Sameer Pradhand, Lance Ramshawd, Nianwen Xued, Ann Taylord, Jeff Kaufmand, Michelle Franchinid, Mohammed El-Bachoutid, Robert Belvind, and Ann Houston. 2013. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia. Linguistic Data Consortium.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Vikas Yadav and Steven Bethard. 2018. 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. Association for Computational Linguistics.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. 2021. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66. Machine Learning and Signal Processing for Big Multimedia Analysis.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230.

A Appendix

A.1 Entity-Level Performance of the Domain Adaptation Model

The entity-level performance, in terms of strict metrics for precision, recall, and F1-score, are presented in Tables 5, 6, 7, and 8, for both Romanian and German languages, on the general and legal domains.

Entity	**Without DA**			**With DA**		
Type	**P**	**R**	**F1**	**P**	**R**	**F1**
DATETIME	55.56	91.01	69.00	46.72	90.07	61.53
EVENT	7.73	55.81	13.58	4.95	49.12	9.00
FACILITY	9.55	69.14	16.78	7.41	73.71	13.47
GPE	58.08	92.94	71.49	49.71	93.40	64.88
LANGUAGE	5.45	87.67	10.26	3.88	87.67	7.44
LOC	17.67	67.46	28.01	13.59	69.05	22.72
MONEY	14.74	87.05	25.21	10.48	83.48	18.62
NAT_REL_POL	38.35	89.76	53.74	30.21	89.24	45.14
NUMERIC	50.22	95.28	65.77	41.44	95.20	57.74
ORDINAL	17.96	80.92	29.39	13.97	85.53	24.02
ORG	40.43	70.44	51.38	38.26	80.69	51.90
PERIOD	12.05	77.11	20.85	9.22	81.00	16.56
PERSON	75.69	89.74	82.12	69.23	90.20	78.34
QUANTITY	17.19	93.90	29.06	12.73	93.90	22.42
WORK_OF_ART	10.96	54.65	18.26	9.04	60.37	15.72

Table 5: Entity-level performance on the RONEC dataset.

Entity	**Without DA**			**With DA**		
Type	**P**	**R**	**F1**	**P**	**R**	**F1**
LEGAL	58.87	83.83	69.17	50.86	80.60	62.37
LOC	46.28	76.57	57.69	41.79	85.30	56.10
ORG	66.13	87.59	75.36	59.88	86.89	70.90
PER	29.37	94.27	44.78	18.17	69.81	28.83
TIME	54.57	90.53	68.09	47.39	91.98	62.55

Table 6: Entity-level performance on the LegalNERo dataset.

Entity	**Without DA**			**With DA**		
Type	**P**	**R**	**F1**	**P**	**R**	**F1**
LOC	69.04	90.97	78.50	70.08	90.33	78.92
LOCderiv	44.32	93.23	60.08	45.97	94.47	61.84
LOCpart	9.34	63.30	16.27	10.30	66.97	17.85
ORG	55.35	80.52	65.60	56.07	79.77	65.85
ORGderiv	0.46	37.50	0.91	0.32	25.00	0.64
ORGpart	17.61	81.40	28.96	17.70	77.91	28.85
OTH	37.86	64.73	47.78	40.18	67.56	50.39
OTHderiv	3.24	56.41	6.13	4.02	66.67	7.59
OTHpart	3.11	50.00	5.85	2.48	38.10	4.66
PER	69.77	94.69	80.34	70.80	94.75	81.04
PERderiv	0.77	45.45	1.51	0.32	18.18	0.64
PERpart	2.81	43.18	5.28	3.74	54.55	7.00

Table 7: Entity-level performance on the GermEval 2014 dataset.

Entity	**Without DA**			**With DA**		
Type	**P**	**R**	**F1**	**P**	**R**	**F1**
AN	4.09	78.26	7.78	6.21	82.61	11.55
EUN	37.84	86.56	52.66	45.91	86.21	59.91
GRT	60.17	98.02	74.57	68.83	98.33	80.98
GS	88.19	98.16	92.91	91.10	97.97	94.41
INN	48.19	90.37	62.86	57.72	91.41	70.76
LD	39.34	97.85	56.12	48.58	97.85	64.92
LDS	6.21	70.00	11.41	9.63	77.50	17.13
LIT	51.27	88.42	64.91	58.22	87.02	69.76
MRK	8.28	76.00	14.93	10.77	68.63	18.62
ORG	30.24	82.05	44.19	37.70	81.55	51.56
PER	42.03	93.96	58.08	51.74	94.26	66.81
RR	20.60	98.20	34.06	27.27	97.30	42.60
RS	82.50	95.20	88.40	85.38	94.31	89.62
ST	22.43	95.31	36.31	28.54	91.41	43.49
STR	5.42	85.71	10.19	7.74	85.71	14.20
UN	32.92	90.21	48.24	40.74	88.94	55.88
VO	22.30	87.94	35.58	29.38	87.94	44.05
VS	16.70	73.55	27.22	21.14	68.00	32.26
VT	52.70	91.36	66.85	60.20	89.52	71.99

Table 8: Entity-level performance on the German LER dataset.

A.2 Comparison with Existing Works

We compare our approach in terms of the strict exact score on each dataset. On the LegalNERo dataset, we extracted the results for the best model (Păiș et al., 2021a) achieving the reported score on the test set. Their approach is similar to ours in that both methods utilize BiLSTM and CRF layers in the architecture. The main difference is that our approach uses BERT embeddings, while Păiș et al. (2021a) generated MARCELL embeddings and employed gazetteers. On RONEC, we considered the results reported for Romanian BERT cased and uncased (Dumitrescu et al., 2020), from their GitHub page[10]. We considered this because our architecture utilizes these pre-trained models as components for embedding generation.

On the German LER dataset, we considered the results for BiLSTM-CRF (Benikova et al., 2014), which utilizes pre-trained embeddings on the German language, and the previously mentioned architecture for predictions. In the end, on the GermEval 2014 dataset, we considered the winning team (Hänig et al., 2014) at the GermEval 2014 competition, which utilizes only the CRF model.

Table 9 showcases the results. We observe that our approach obtains comparable results on LegalNERo; on RONEC and German LER, the differ-

[10] https://github.com/dumitrescustefan/ronec/tree/master/evaluate

Method	LegalNERo	RONEC	LER	GermEval 2014
MARCELL+BiLSTM+CRF (Păiș et al., 2021a)	85.34	-	-	-
romanian-bert-cased (Dumitrescu et al., 2020)	-	91.9	-	-
romanian-bert-uncased (Dumitrescu et al., 2020)	-	95.2	-	-
BiLSTM-CRF (Benikova et al., 2014)	-	-	95.46	-
CRF (Hänig et al., 2014)	-	-	-	79.08
Our method	85.17	87.5	92.30	85.75

Table 9: Comparison with existing works. All scores are F1-strict scores, in percentages (%).

ence between 4% and 8%, and on the GermEval 2014 dataset, our method performs better than a CRF model by 6%.

A.3 Embedding Space Visualization

We analyze the embedding space by employing t-SNE representations on the embedding space generated with the pre-trained BERT models (before fine-tuning). Since some named entities may have more than one word or token, we average embeddings to generate a meaningful representation. Formally, given the set of token embeddings w_i^j, each token of the Transformer's input has the following embedding representation e^j:

$$e^j = \frac{1}{N^j} \sum_{i=1}^{N^j} w_i^j \tag{15}$$

where N_j represents the length of the jth named entity.

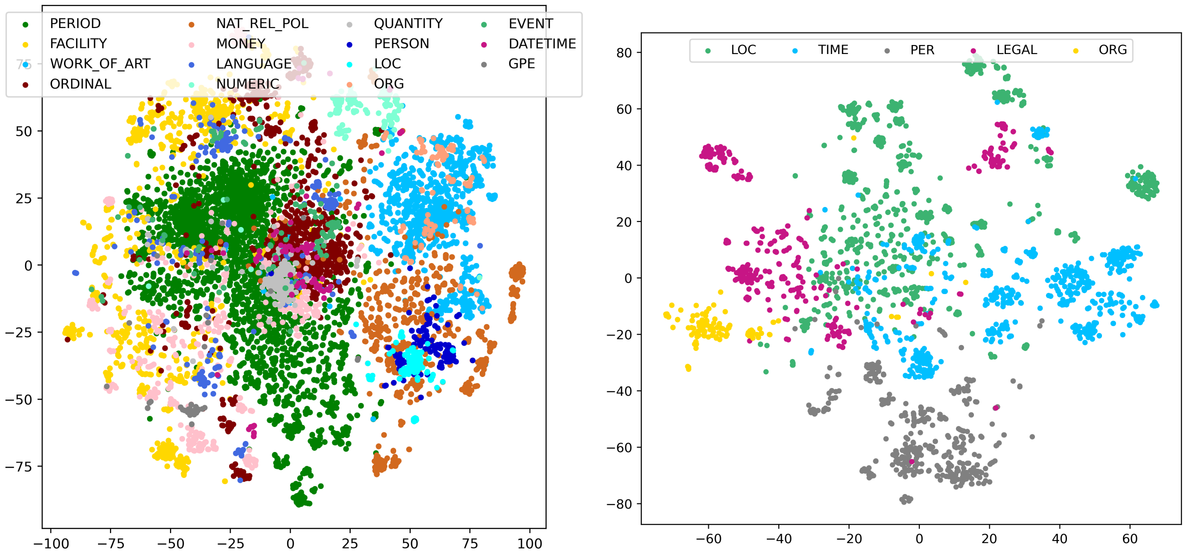
In this space, we apply t-SNE to generate the visualizations on the test set of each dataset we utilized in this work, the perplexity being set to 30. Figure 4 shows the plots on the Romanian language, while Figure 5 presents the visualizations on the German language. We observe that tokens from the same class cluster together in both languages. In addition, we can observe that LegalNERo is the sparsest dataset, with classes in general well separated. On the other hand, we see the data's tendency to cluster together on the LER dataset, but compared to the general domain, it is a less linearly separable dataset. However, the models trained on this dataset obtained better results due to model over-parametrization. We see linearly separable clusters in the t-SNE representations on the general domains, with some scattered points.

A.4 Case Study

We present examples of the outputs produced by the domain adaptation model in Table 10 from Appendix A.5.

In the case of the Romanian language, we see that the boundaries are not well recognized, such as in *"Trezoreria Statului"* (eng. "State Treasury"), where only *"Trezoreria"* is marked as an organization. In other cases, such as *"Băncii Naționale a României,"* (eng. "of the Romanian National Bank"), the comma is included in the entity. Other limitations rely on the misclassification of entity type, identifying entities that are not annotated in the ground truth, such as locations, dates, and organizations (although these can be considered entities in other contexts or can be subject to the difficulty for annotating datasets and ambiguity of words), and not identifying entities if are used in different contexts in the same sentence (for example, words that possess an indefinite/definite article; e.g., in the RONEC dataset, we have *"persoane fizice"* (eng. "natural persons") with both words annotated or only *"persoane"* (eng. "persons")). The model does not capture this variety in the dataset, which overfits this scenario. As presented in the Subsection 5.1, some entities are misclassified with other similar types, such as an event with an organization, when they present acronyms (for example, *"USFL"* - United States Football League and *"NFL"* - National Football League) and can be used interchangeably.

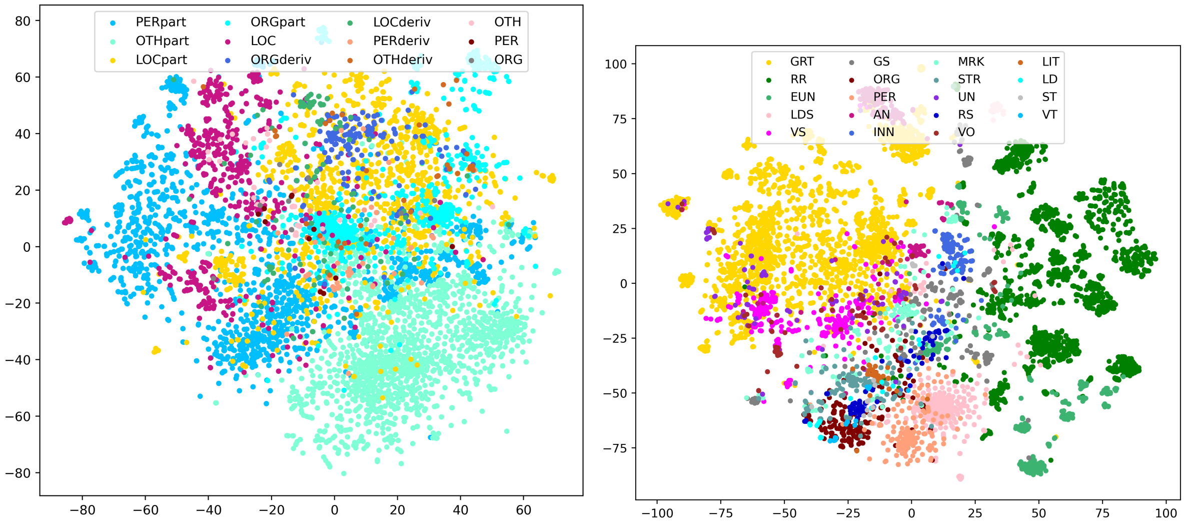
In the case of the German language, the model predicts the wrong boundaries rather than identifying the entity class (this is also supported by the higher partial metric than the strict and exact metrics). One such example can be seen in Table 10 on the GermEval 2014 dataset, where *"Przemyslaw II. von Großpolen"* (which is a name, where *"von Großpolen"* means "from Greater Poland" in English) is identified as two entities, namely the primary name *"Przemyslaw II."* and the location *"Großpolen"* which is from the name. Another limitation, which is not present in the Romanian dataset, is the identification of long entities. For example, *"Stellungnahme des Wissenschaftlichen Beirats beim Bundesministerium der Finanzen aus*



(a) RONEC dataset

(b) LegalNERo dataset

Figure 4: t-SNE visualizations of the embedding space on Romanian datasets.



(a) GermEval 2014 dataset

(b) German LER dataset

Figure 5: t-SNE visualizations of the embedding space on German datasets.

dem Jahr 2010, Reform der Grundsteuer, S. 6" (eng., "Statement of the Scientific Advisory Board at the Federal Ministry of Finance from 2010, reform of the property tax, P. 6") which is not identified by our system. Finally, the last limitation is that the model does not identify the correct entity types. It can be viewed in Table 10, under the LER dataset, where *"A"* is a placeholder for a person, and *"X"* is a placeholder for a city. These were identified as company and location, respectively, just from the context. In this instance, we shall recall that LER has 19 fine-grained classes, some corresponding to coarse-grained, higher-level classes.

A.5 Example of Predictions for the Domain Adaptation Model

	LegalNERo
GT.	Punerea în circulație a monedelor de circulație, cu tema Vizita Apostolică a Sanctității Sale Papa Francisc **PER** în România **LOC** , se va face prin sucursalele regionale București **LOC** , Cluj **LOC** , Iași **LOC** și Timiș **LOC** ale Băncii Naționale a României **ORG** , cu ocazia efectuării plăților în numerar către instituțiile de credit / Trezoreria Statului **ORG** .
Pred.	Punerea în circulație a monedelor de circulație, cu tema Vizita Apostolică a Sanctității Sale Papa **PER** Francisc **PER** în România **LOC** , se va face prin sucursalele regionale București **LOC** , Cluj **LOC** , Iași **LOC** și Timiș **LOC** ale Băncii Naționale a României **ORG** , cu ocazia efectuării plăților în numerar către instituțiile de credit / Trezoreria Statului. **ORG**
	RONEC
GT.	Această regulă , reglementată în prezent la art. 83 **NUMERIC** alin. (3 **NUMERIC**) din Codul de procedură fiscală, republicat în M.O. nr. 863 **NUMERIC** / 26.09.2005 **DATETIME** , a generat unele efecte în sensul că în practică au existat numeroase situații în care suma impozitului datorat depășea suma venitului (de exemplu există foarte mulți acționari **PERSON** persoane fizice **PERSON** cu dividende sub un leu nou **MONEY**).
Pred.	Această regulă , reglementată în prezent la art. 83 **NUMERIC** alin. (3 **NUMERIC**) din Codul de procedură fiscală, republicat în M.O. nr. 863 **NUMERIC** / 26.09.2005 **DATETIME** , a generat unele efecte în sensul că în practică au existat numeroase situații în care suma impozitului datorat depășea suma venitului (de exemplu există foarte mulți acționari **PERSON** persoane fizice cu dividende sub un leu nou).
	GermEval 2014
GT.	Mit Herzog Przemysław II. **PER** von Großpolen **LOC** schloss Mestwin **PER** am 15. Februar 1282 im Vertrag von Kempen **LOC** eine „donatio inter vivos“ (Geschenk unter Lebenden) und vermachte ihm sein Herzogtum.
Pred.	Mit Herzog Przemysław II. von Großpolen **PER** schloss Mestwin **PER** am 15. Februar 1282 im Vertrag von Kempen **LOC** eine „donatio inter vivos“ (Geschenk unter Lebenden) und vermachte ihm sein Herzogtum.
	LER
GT.	Sie hatte in den Streitjahren bei der A **PER** mit Sitz in X **ST** (Österreich **LD**) Reisevorleistungen zur Durchführung von in der Bundesrepublik Deutschland **LD** (Deutschland **LD**) ausgeführten Radtouren bezogen.
Pred.	Sie hatte in den Streitjahren bei der A **UN** mit Sitz in X **ST** (Österreich **LD**) Reisevorleistungen zur Durchführung von in der Bundesrepublik Deutschland **LD** (Deutschland **LD**) ausgeführten Radtouren bezogen.

Table 10: Examples of ground truth (GT.) labels and predictions (Pred.) for the domain adaptation models. We selected the examples that have wrong predictions. Best viewed in color.