# Did You Enjoy the Last Supper? An Experimental Study on Cross-Domain NER Models for the Art Domain

**Alejandro Sierra-Múnera**
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
alejandro.sierra@hpi.de

**Ralf Krestel**
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
ralf.krestel@hpi.de

## Abstract

Named entity recognition (NER) is an important task that constitutes the basis for multiple downstream natural language processing tasks. Traditional machine learning approaches for NER rely on annotated corpora. However, these are only largely available for standard domains, e.g., news articles. Domain-specific NER often lacks annotated training data and therefore two options are of interest: expensive manual annotations or transfer learning.

In this paper, we study a selection of cross-domain NER models and evaluate them for use in the art domain, particularly for recognizing artwork titles in digitized art-historic documents. For the evaluation of the models, we employ a variety of source domain datasets and analyze how each source domain dataset impacts the performance of the different models for our target domain. Additionally, we analyze the impact of the source domain's entity types, looking for a better understanding of how the transfer learning models adapt different source entity types into our target entity types.

## 1 Introduction

Cultural heritage archives contain vast amounts of unstructured data where valuable knowledge resides. This data can be analyzed and valuable information can be extracted using natural language processing (NLP) tools. Nowadays, most of the NLP tasks are performed using deep learning models which rely on large amounts of training data.

One of the core NLP tasks is named entity recognition (NER) which consists of finding mentions of *named entities* from a usually pre-defined set of entity types. Machine learning models learn entity and context patterns from labeled corpora allowing them to discover new entity mentions from unseen text.

In a scenario where there is a previously annotated large corpus, these models achieve good performance and can find new named entities from the pre-defined set of entity types. In the past, such datasets have been built for domains such as news wire (Tjong Kim Sang and De Meulder, 2003) and biomedical texts (Stubbs and Uzuner, 2015) containing annotations for entity types, such as *person*, *location*, *organization*, or *protein*, and *gene expression*.

Large labeled corpora are expensive and time-consuming to obtain. For less popular domains, large annotated corpora typically don't exist. There is especially a lack of annotated data for domain-specific entity types. One specific domain that requires non-standard entity types to be extracted is the cultural heritage domain. In this paper, we focus on digitized art-historic archives, in particular on the entity type *artwork*. For this particular entity type, there are no extensive datasets. The entity type *artwork* is different from the standard ones (person, location, organization, date) and poses some interesting challenges (Jain and Krestel, 2019). Not only is the entity type different, but also the structure and noise of digitized art-historic texts are different from news wire or biomedical text collections.

One particular challenge is the ambiguity inherent to the definition of such titles due to the fact that sometimes these titles describe a scene or contain other named entities. For instance, the painting titled '*Girl before a mirror*' by Pablo Picasso, depicts a girl before a mirror. Only the context of the mention identifies this phrase as a painting title.

Moreover, a big percentage of art-historic archives need to be digitized first using optical character recognition (OCR) software. This routinely introduces errors such as mis-identified characters, the addition of noise, and the loss of formatting structure(van Strien et al., 2020; Lin, 2003; Ro-



*Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, pages 173–182
co-located with ICON-2021, December 16-19, 2021, Silchar, India. 
https://doi.org/10.26615/978-952-94-5833-2_020

driquez et al., 2012). Further, the quality of OCRed texts strongly depends on the print quality of the original documents (Traub et al., 2015; Rodriquez et al., 2012; Mieskes and Schmunk, 2019).

Given the aforementioned challenges, different alternatives for solving the task could be explored. One would be manually annotating a large corpus with artwork title information. But besides being a time-consuming and expensive task, it would not scale to further cultural heritage entities such as galleries, art styles, or art movements. Another would be focusing on gazetteers and rule-based approaches. But, listing all possible artwork titles would not only be cumbersome, but would also not solve the ambiguity problem for phrases such as '*Girl before a mirror*'.

The most promising approach is to make use of existing, previously annotated corpora from other domains and *transfer* the learned patterns to the new domain. In combination with deep learning models, this domain adaptation via *transfer learning* or *multi-task learning* has shown good results for popular domains (Rodriguez et al., 2018). Different models have been proposed in the past under the concept of *cross-domain NER* to solve the problem. These models learn to identify named entities within a *target domain* based on patterns learned from a large, labeled dataset from a *source domain*.

The goal of this paper is to evaluate the performance of some of the best of those models for the artwork recognition task. The paper is structured as follows: In Section 2 we describe the existing cross-domain NER models. In Section 3 we describe the existing NER datasets available for different domains and the construction of a target dataset used for the training and evaluation of artwork recognition. In Section 4 we describe the evaluation setup and in Section 5 the results are outlined and finally, in Section 6, the conclusion and future work are proposed.

## 2 Related Work

In this section, we discuss previously proposed cross-domain NER models and focus on the domain adaptations that those models propose.

Cross-domain NER models could be divided into two main categories. The models in which the source and target domain share the entity types but have differences in terms of the vocabulary, and the models which consider the disparity between the entity types in the source and the target domain.

For instance, one traditional task for the first group would be the transfer of *persons*, *locations* and *organization* from the news domain into the social media domain. In this case, the persons mentioned in the news domain might be different from the persons mentioned in social media. Moreover, the language used in social media is different from the language used in news articles. Artifacts, such as emojis, hashtags, or '@' as well as the structure of sentences differ between domains. However, the entity types remain constant in the domain adaptation task. Liu et al. (2020a) propose a model for low-resource target domains combining multi-task learning (MTL) and a mixture of entity experts (MoEE), aiming to improve generalization and reduce the over-fitting effect when a model learns entities from a source domain. Zhou et al. (2019) propose a general neural transfer framework called Dual Adversarial Transfer Network (DATNet), Wang et al. (2020) extend the popular Bi-LSTM-CRF architecture for multi-domain NER, dividing the domain-specific and independent components of the network, to achieve adaptation over multiple genres.

Other models deal with different entity types in the target domain compared with the source domain. This is the case for domain-specific entity types where extensively annotated corpora are missing. Artwork mentions, for instance, are not annotated in traditional NER datasets, therefore we focus our study on this kind of domain adaptation. Within these models, Lee et al. (2018) proposed to transfer the weights of a Bi-LSTM-CRF model with both word and character embeddings. The weights were trained on the source domain and then fine-tuned on the smaller target dataset. They experimented with transferring different parts of the network to the target domain and concluded that transferring the weights from the lower layers of the network, particularly the character Bi-LSTM layer, improved the performance of the NER model on the target domain compared to a model trained only with the target dataset (no transfer). A similar model proposed by Lin and Lu (2018), called CDMA-NER augmented the idea of using a pre-trained model by including adaptation layers on top of it to perform the domain adaptation without the need of retraining the source model. The adaptation between domains is based on the bottom layer of the Bi-LSTM model, particularly on the adaptation of word embeddings.

A different kind of domain adaptation is used by models which simultaneously train the source and the target domain in a multi-task learning approach (Bhatia et al., 2018), (Beryozkin et al., 2019), (Jia and Zhang, 2020). In the multi-task model proposed by Jia and Zhang (2020) (Multi-Cell Compositional LSTM for NER Domain Adaptation) which is based on an LSTM network, each entity type has an independent cell state. Additionally a compositional cell combines all the entity type cells into the final output which is then passed to the domain-specific conditional random field (CRF). The domain adaptation is performed on the entity type level, and the model leverages the context embeddings provided by BERT (Devlin et al., 2019). In their experiments, they transferred information from the news domain to the biomedical and the social media domain.

Another recently proposed model called CrossNER (Liu et al., 2021) introduces domain-adaptive pre-training (DAPT) as a technique to continue the pre-training of language models such as BERT with domain-specific raw texts by masking spans of tokens instead of random tokens for training. They experimented with different masking strategies as well as different corpora selection criteria and concluded that the best performance is obtained when DAPT is performed in a set of sentences containing general and task specific entities. The entities used for corpora selection are chosen from predefined resources, such as gazetteers or knowledge graphs. Besides the pre-trained language model, which is trained in the target domain, the model uses a linear layer on top. They experimented with training the whole model on only the target domain, jointly training the source and target domain, and pre-traning in the source domain followed by fine-tuning on the target domain. Their results show that pre-training followed by fine-tuning yields better results. In their paper, they also introduce a new dataset with a diverse set of annotated texts from different domains with domain-specific entity types. In their experiments these domains are treated as targets. We use their dataset for our experiments but instead of treating them as target domains, we consider them as source domains. The details of the datasets we use are described in Section 3.

## 3 Datasets

As mentioned in Section 2, cross-domain NER relies on the knowledge of a source domain. This knowledge can be in the form of an annotated corpus with domain-specific entity types or a pre-trained model specialized in recognizing them. We use a diverse set of source datasets for this purpose. Regarding the target domain, we created a dataset with sentences containing art-related entity mentions. In the following subsections, we describe the datasets considered in our experiments as source datasets, as well as the target domain dataset which will serve as a training, validation, and test dataset.

### 3.1 Source Datasets

For source domain datasets, we consider the widely used CoNLL03 (Tjong Kim Sang and De Meulder, 2003) English dataset which consists of news texts annotated with the traditional named entity types: *person*, *location* and *organization* plus *miscellaneous*.

To study the impact of the source domain and its entity types, we also consider the dataset published by Liu et al. (2021): a collection of manually annotated corpora from five domains (artificial intelligence, music, literature, politics and science) that was labeled with domain-specific entity types. The variety in entity types is important in our evaluation because we focus on domain adaptation approaches that specifically need to deal with different entity types. In their paper, (Liu et al., 2021) used the newly labeled corpora as target domains, and the goal was to perform domain adaptation from the news domain to these, therefore the target training set was smaller than the validation and the test set, thus limiting the amount of labeled data in the target domain. In our experiment we consider those datasets as source datasets, therefore we split the corpora in a different way to increase the size of the training set.

Also these datasets contain only sentences mentioning at least one entity. Therefore, to have a fair comparison between source datasets, we filter the CoNLL03 dataset to keep only the sentences that mention an entity, we refer to the filtered dataset as the news dataset. This comprises the following reduction of sentences for the news dataset: the training set is reduced from 14,041 to 11,132 sentences, and the validation set is reduced from 3,250 to 2,605 sentences.

The resulting group of source datasets is referred to in our experiments as the unbalanced source datasets, due to the difference in sizes.

Additionally, in order to compare the impact of

| Dataset | Balanced | | Unbalanced | |
|---|---|---|---|---|
| | Train | Val. | Train | Val. |
| News | 781 | 100 | 11132 | 2605 |
| AI | 781 | 100 | 781 | 100 |
| Literature | 781 | 100 | 816 | 100 |
| Music | 781 | 100 | 845 | 100 |
| Politics | 781 | 100 | 1192 | 200 |
| Science | 781 | 100 | 993 | 200 |

Table 1: Number of Sentences in Source Domain Datasets

| | Train | Val. | Test |
|---|---|---|---|
| Sentences | 180 | 70 | 294 |
| Mentions | 51 | 21 | 74 |

Table 2: Art Target Domain Dataset

the source domains and avoiding the possible bias of the dataset sizes, we under-sample each of the datasets except the AI dataset, being the smallest, to generate source datasets with exactly the same number of sentences in the training and validation sets. The resulting sizes in terms of sentences in the training and validation sets are detailed in Table 1.

### 3.2 Target Dataset

Our study focuses on the detection of artwork mentions in digitized art-historic documents. However, there is no public dataset available with artwork title annotations. Therefore, we manually annotated a set of randomly extracted 544 sentences for the evaluation of the different models. An annotation tool was used by two non-expert annotators, and afterwards the inter-annotator agreement in the results was analyzed. The Fleis-kappa (Fleiss, 1971) value was $-1.86$ and Krippendorf-alpha (Krippendorff, 1970) $0.61$ meaning that there was poor agreement among annotators. As expected, even for humans, the annotation process was difficult due to the challenges expressed in Section 1. Therefore, an additional step of manual revision of each annotation was performed, and the disagreements were resolved with the help of web search. Afterwards, the target domain dataset was split into train, validation and test, with the sizes depicted in Table 2.

## 4 Experimental Setup

Our evaluation aims to shed light on the power of different cross-domain NER models to adapt to the art domain and recognize artwork mentions. For our experiments, we focus on the models **CDMA-NER** proposed by Lin and Lu (2018), **Muli-Cell LSTM** proposed by (Jia and Zhang, 2020) and **CrossNER** proposed by Liu et al. (2020b). To compare the performance, we train each of these models using datasets from a set of source domains and a single target domain training dataset. We measure the F1 score for the task of recognizing artwork mentions in the target test set.

For the experiments with CDMA-NER, GloVe (Pennington et al., 2014) word embeddings are used, and for both, Multi-Cell LSTM and CrossNER, which are designed to use pre-trained language models, we use BERT base model (cased) as well as an adaptation to the art domain following the domain-adaptive pre-training used in CrossNER.

### 4.1 Domain-Adaptive Pre-Training

We further pre-train the BERT base model (cased) for Multi-Cell LSTM and CrossNER with a set of raw art-related texts, we generated a set of 500,000 sentences extracted from digitized art-historic documents containing artwork titles from the Getty vocabularies (Harpring, 2010). Specifically, we perform a string match of sentences against the Cultural Objects Named Authority (CONA) vocabulary[1] and the Union List of Artist Names (ULAN)[2] containing titles of artwork and architecture, and artist names, respectively. With the 500,000 sentences, pre-training is performed for 15 epochs as proposed by Liu et al. (2020b).

### 4.2 Training

Each model was trained for a maximum of 500 epochs with early stopping and the validation set was used to determine when the model did not need further training and the best model was evaluated against the target test dataset. In the case of CDMA-NER and CrossNER, the source validation dataset was used to determine the best source model, before transferring the weights to the target domain

[1]Getty CONA (2017), http://www.getty.edu/research/tools/vocabularies/cona, accessed October 2021.

[2]Getty ULAN (2017), http://www.getty.edu/research/tools/vocabularies/ulan, accessed October 2021.

| | |
|---|---|
| News | [Liam Gallagher:person] , singer of [Britain:location] 's top rock group [Oasis:organization] , flew out on Thursday to join the band three days after the start of its [U.S.:location] tour |
| AI | Examples of [supervised learning:field] are [Naive Bayes classifier:algorithm] , [Support vector machine:algorithm] , [mixtures of Gaussians:algorithm] , and network |
| Literature | It tied with [Roger Zelazny:writer] ' s [This Immortal:book] for the [Hugo Award:award] in 1966 |
| Music | Two of his most popular recordings were [Layla:song] , recorded with [Derek and the Dominos:band] ; and [Robert Johnson:musical artist] ' s [Cross Road Blues:song] , recorded with [Cream:band] |
| Politics | Three [United States:country] presidents have been impeached by the [House of Representatives:misc] : [Andrew Johnson:politician] in 1868 , [Bill Clinton:politician] in 1998 , and [Donald Trump:politician] in 2019 . |
| Science | The journal establishment was similar to the starting of [The Astrophysical Journal:academic journal] and [The Astronomical Journal:academic journal] by [George Ellery Hale:scientist] |
| Art | Figure 39 . [On the Terrace:artwork] , 1867 . Panel , 17.7 x 18 cm . © The Cleveland Museum of Art , Bequest of Clara Louise Gehring Bickford , 1986.68 . Photo : Courtesy of the Museum . |

Table 3: Dataset Examples

training. For all models their publicly available implementations were adapted to use the dataset configuration proposed in this paper.

Additionally, a baseline model was trained without using source domain data. This baseline model is based on the Bi-LSTM-CRF model originally proposed by Lample et al. (2016), and implemented using FlairNLP (Akbik et al., 2019). It was trained ten times using BERT base (cased) as the embedding model, the average F1 over the 10 runs is reported in Table 4.

The under-sampling process to generate the size-balanced source datasets is repeated 5 times to generate random subsets of the data. For the smaller AI source dataset, 5 shuffled versions with the same sentences are used to train the models. The results for the size-balanced experiments in Table 4 show the average performance over the 5 runs.

## 5 Results

Table 4 shows an overview of the results in terms of F1-measure for each of the evaluated models using the different source datasets, plus the performance of the baseline model. The first observation is that the baseline model achieves very competitive results in comparison to the cross-domain models. In only one occasion the other models were able to outperform the baseline, which suggests that the transfer learning approach seems to work in very specific settings.

Another observation is that DAPT is in general not improving the language model for the CrossNER model, which performs better with the original BERT model. One possible reason is the digitization noise introduced into the raw text used to perform DAPT. For Multi-cell LSTM, the average improvement is very small. The CDMA-NER model in general performs worse than the other models and the baseline, and the reason could be the lack of contextualized representation of words in the GloVe embeddings.

Generally, the CrossNER model performs better than the other two models and its performance is similar to the baseline, although the model is relatively simple in comparison to Multi-Cell. This suggests that the traditional LSTM-CRF combination might not be suitable for transfer learning to complex entities such as artworks. The combination of LSTM and CRF is positive for NER as shown by the performance of the baseline model, but as the architecture becomes more complex, the performance is compromised. Another reason why Multi-

| **Baseline**: FlairNLP$_{BERT}$ | .589 | | | | | | |
|---|---|---|---|---|---|---|---|
| **Cross-Domain NER Models** | **Source Domain** | | | | | | |
| | **News** | **AI** | **Lit** | **Mus** | **Pol** | **Sci** | **Avg** |
| CDMA-NER | .460 | .368 | .344 | .394 | .409 | .413 | .386 |
| Multi-cell LSTM$_{BERT}$ | .255 | .509 | .385 | .467 | .438 | .459 | .451 |
| Multi-cell LSTM$_{DAPT}$ | .343 | .487 | .436 | .464 | .471 | .413 | .454 |
| CrossNER$_{BERT}$ | .537 | .519 | .578 | .535 | .521 | .512 | .533 |
| CrossNER$_{DAPT}$ | **.594** | .488 | .507 | .477 | .482 | .528 | .496 |
| | Size-balanced experiments | | | | | | |
| CDMA-NER | .332 | .339 | .365 | .336 | .360 | .318 | .344 |
| Multi-cell LSTM$_{BERT}$ | .495 | .455 | .460 | .446 | .484 | .441 | .457 |
| Multi-cell LSTM$_{DAPT}$ | .434 | .454 | .489 | .458 | .415 | .463 | .456 |
| CrossNER$_{BERT}$ | .522 | .535 | .518 | .543 | .517 | .586 | .540 |
| CrossNER$_{DAPT}$ | .475 | .503 | .516 | .560 | .528 | .518 | .525 |

*The results in bold font correspond to values higher than the baseline*

Table 4: F1-Scores for Art Target Domain

Cell LSTM models might be performing worse than CrossNER is the fact that there is no overlap between source and target entity types, therefore the weights within the LSTM cells are not being strongly shared among domains.

The results of training the models with the unbalanced datasets reveal that the size of the source dataset does not guarantee a good target performance. The adapted news dataset is 13 times bigger than the music and literature datasets, but the performance is comparable when training the CrossNER$_{BERT}$ model. One reason for this behavior is the more general definition for entity types in CoNLL03, different from the more specialized entity types in the music and literature datasets.

One of the aspects which differentiate the various domains is the set of entity types that are relevant for the domain and are present in the different datasets. To study the impact on the performance of artwork recognition we remove individual entity types from the full music dataset. For each of the 13 entity types in this dataset, we generate an alternative version of the dataset in which the entity type is not considered in the annotations. This means that the tokens which were previously labeled as part of those named entities will remain in the dataset but without the annotation. Each altered dataset is used to train the 5 studied models. In Figure 1, the models' performance after altering the dataset is displayed as relative performance change with respect to the original experiment with the complete dataset. This way, we intend to analyze how each model depends on the source entity types to be able to transfer that knowledge to the recognition of artwork mentions.

From the figure it is clear that the Multi-cell LSTM model suffers a greater decrease in performance when the musical artists and bands are not present in the source dataset. This is an indicator of the manner in which this model learns the connections between the source and target entity types through the entity-typed LSTM cells. It is interesting, however, that in some cases the performance improves when removing entity types. This suggests that the model is sensitive to the similarity between the source and target entity types. Thus, depending on the type of entities we would like to recognize in the target domain, we should select the source dataset. Best results are achieved with the most similar entity types in the source domains. To phrase it in terms of the artwork recognition task, it would make sense to first analyze which domains contain titles of human-created creative works and then use those entity types exclusively.

Figure 2 depicts results of a similar experiment. In this case only one of the entity types is present in the dataset. Comparing both figures, it is clear that source datasets with just one entity type perform worse than source datasets with more variety in entity types. It is, however, counter-intuitive that the entity types which help the most in the transfer setting towards recognizing artworks are not

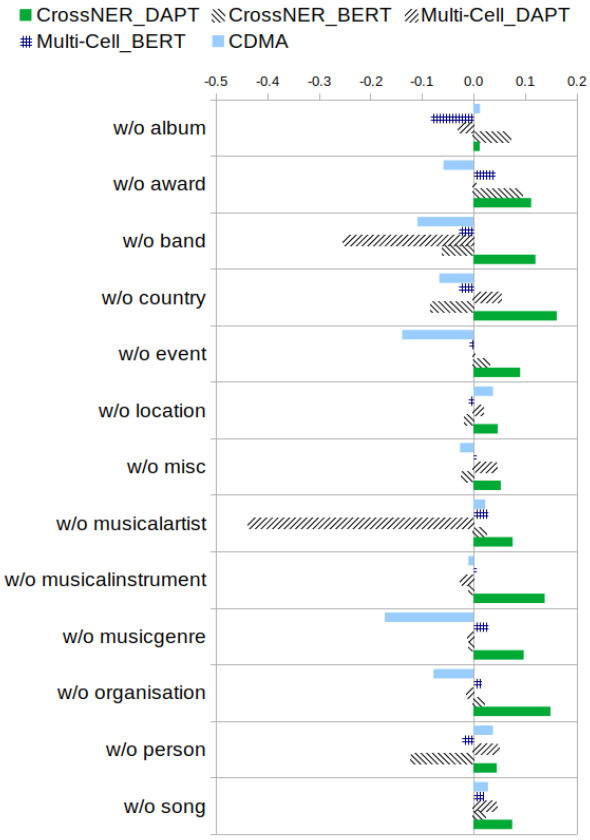


Figure 1: Change in F1 score when one entity type is removed from the source dataset *music*

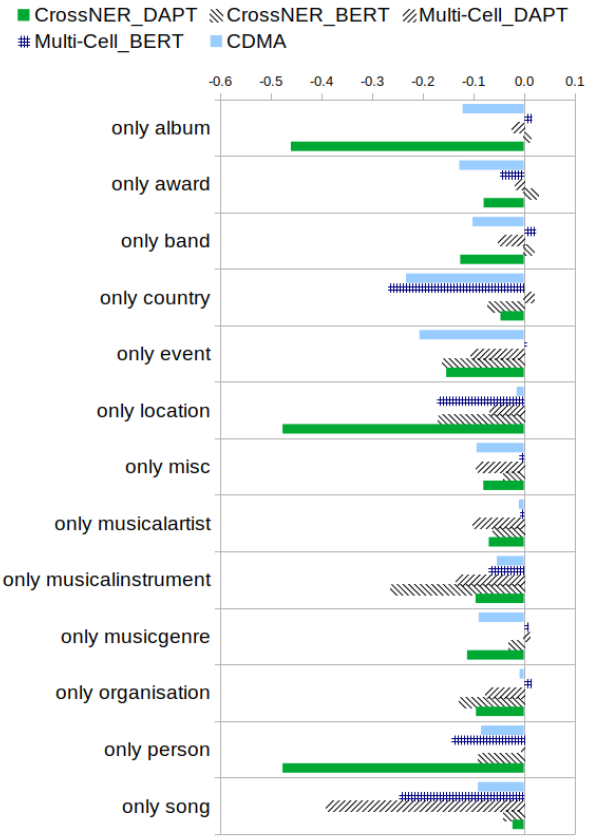


Figure 2: Change in F1 score when only using one entity type in the source dataset *music*

*song* or *album*, which are the entity types in the music domain that resemble closest to the notion of artwork.

Additional details of the results can be found in https://github.com/HPI-Information-Systems/cross-domain-ner

### 5.1 Qualitative Analysis

Besides the quantitative evaluation, we also performed an error analysis by investigating example predictions of the models. Specifically, we analyse the models trained with the original *music* source dataset.

Firstly, Table 5 example E1 shows a sentence which contains a correctly recognized entity mention and a typical error in which the model is able to recognize the presence of an artwork but the boundaries are not correctly identified. For other models the determiner of the second mention was part of the title, which is not the case in this particular example E1, but it is a persistent error for all models. The presence of the article in the titles is a complex boundary to define even for humans since there is no clear rule that could be applied.

In example E2, we see a false positive predicted by CDMA-NER and not predicted by any other model. One possible reason for this error is the lack of context in the sentence, the presence of a name at the beginning, and the quotation marks. Without the knowledge that Claude Monet is a painter, it would be hard to distinguish it from an artwork mention, given that many paintings are named after persons.

In example E3, the sentence is particularly long and contains many artwork mentions. The CrossNER$_{BERT}$ model, which is the best performing one, is able to identify all the mentions but fails to set the correct initial boundaries for three. One specially interesting observation is that 2 titles follow the pattern *'Painter* {*WORD*} *His* {*WORD*}*'* but the model is able to correctly recognize only one of them.

The fourth example E4 exemplifies a very challenging artwork title to recognize. It is a notably long title containing a combination of uppercase and lower case words and references to different locations. In our experiments, no model was able to recognize the artwork mention in that sentence.

| | |
|---|---|
| E1 | ... as in [The Confidence] ( Salon of 1857 , Fig . 26 ) . Even relaxed in a tavern , as in the [Smoker in Black] ( Fig . 27 ) , the ... (Predicted by Multi-Cell LSTM$_{DAPT}$) |
| E2 | 19Robinson , " Claude Monet , " 698 (Predicted by CDMA-NER) |
| E3 | The autobiographical dimension is furthered by Meissonier 's inclusion of works he had created or owned . [Painter Showing His Drawings] , set in the quai Bourbon studio,3 includes an enlarged [Samson Battling the Philistines] , perhaps hinting that the artist is similarly an inspired hero , coping with his own philistine world . Identifiable below is the unframed [Smoker] of 1842 ; in the center , [The Evangelists] , which is propped against a portrait recognizably of Meissonier ; and in the portfolio , a drawing for [The Evangelists] ( Musée du Louvre , rf 1908 ) . In the background of [Painting Collectors] is also an enlarged version of the [Martyrdom of Saint Lawrence] used in the [Painter in His Studio] of 1843 and an Italian - school painting of a half - length , seminude woman that belonged to Meissonier.4 (Predicted by CrossNER$_{BERT}$) |
| E4 | We are grateful to Cecilia Powell for pointing out the conflation in Wilton , op . cit . , of this watercolor with the [View down the Mosel from the Hillside above Pallien] ( circa 1839 , illustrated in Powell , op . cit . , p. 132 ) , and to Peter Bower for his assistance in preparing this catalogue entry . (All models failed to recognize the mention) |
| E5 | ... lent from the distinguished collection of Mrs Walter Jones , the widow of Walter H. Jones . Her other loans included the [Red Rigi] ( no . 891 ) , the [Blue Rigi] ( no . 895 ) , [Venice , Mouth of the Grand Canal] ( no . 899 ) and [Mainz and Castel] ( no . 904 ) . When the drawing was sold ... (Predicted by CDMA-NER) |

*Squared brackets represent ground truth and highlighted text represents predicted annotations*

Table 5: Inference Examples

## 6 Conclusions

In this paper, we studied the task of complex NER, specifically recognizing artworks in art-historic texts. We discuss the reasons why this is a hard task and why it is promising to leverage annotations from other domains to compensate for the lack of annotated resources for the art domain. We explained the concept of cross-domain NER using transfer learning which has been investigated in the past to achieve the aforementioned domain adaptation and presented related work connected to this concept. Based on the problem setup and a collection of annotated datasets, we performed a set of experiments to understand the performance of domain-adapted NER to recognize artworks. In the experiments we analyzed both, the models and the datasets, in order to isolate and understand independently different aspects of the presented approaches. From the experimental evaluation of Cross-domain NER approaches for the recognition of artworks we conclude that, although domain adaptation is a promising approach to achieve this goal, a simpler alternative, namely a LSTM-CRF model with BERT base (cased), perform as well as the best Cross-domain NER.

As future work, we would like to investigate the explainability and interpretability of cross-domain NER models to understand better their limitations and propose new models that not only take into account the differences in terms of entity types and language between domains, but also semantic relations between the domains and the named entities. Additionally, it would be of interest to investigate the domain adaptation of other tasks like information extraction and knowledge graph embedding models, which could be jointly trained with NER.

## Acknowledgements

We thank the Wildenstein Plattner Institute for providing the digitized corpus used in this work. This research was funded by the HPI Research School on Data Science and Engineering.

## References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages

54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Genady Beryozkin, Yoel Drori, Oren Gilon, Tzvika Hartman, and Idan Szpektor. 2019. A joint named-entity recognizer for heterogeneous tag-sets using a tag hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 140–150, Florence, Italy. Association for Computational Linguistics.

Parminder Bhatia, Kristjan Arumae, and Busra Celikkaya. 2018. Dynamic transfer learning for named entity recognition. *Studies in Computational Intelligence*, 843:69–81.

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.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Patricia Harpring. 2010. Development of the Getty vocabularies: AAT, TGN, ULAN, and CONA. *Art Documentation: Journal of the Art Libraries Society of North America*, 29(1):67–72.

Nitisha Jain and Ralf Krestel. 2019. Who is mona l.? identifying mentions of artworks in historical archives. In *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, Proceedings*, volume 11799 of *Lecture Notes in Computer Science*, pages 115–122. Springer.

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, Online. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

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.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Xiaofan Lin. 2003. Impact of imperfect ocr on part-of-speech tagging. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 284–288 vol.1.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020a. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 1–6, Online. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Peng Xu, Zhaojiang Lin, and Pascale Fung. 2020b. Cross-lingual spoken language understanding with regularized representation alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7241–7251, Online. Association for Computational Linguistics.

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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13452–13460. AAAI Press.

Margot Mieskes and Stefan Schmunk. 2019. OCR quality and NLP preprocessing. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 102–105, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Juan Diego Rodriguez, Adam Caldwell, and Alexander Liu. 2018. Transfer learning for entity recognition of novel classes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1974–1985, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw OCR text. In *Konvens*, pages 410–414.

Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*, pages 484–496. SCITEPRESS.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *J. Biomed. Informatics*, 58:S20–S29.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Myriam C Traub, Jacco Van Ossenbruggen, and Lynda Hardman. 2015. Impact analysis of ocr quality on research tasks in digital archives. In *International Conference on Theory and Practice of Digital Libraries*, pages 252–263. Springer.

Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8476–8488, Online. Association for Computational Linguistics.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy. Association for Computational Linguistics.