Overcoming Shortage of Training Data for Cultural Heritage Analysis

Nitisha Jain *, Ralf Krestel
Information Systems Group, Hasso-Plattner-Institute, University of Potsdam, Germany
E-mails: nitisha.jain@hpi.de, ralf.krestel@hpi.de

Abstract. As machine learning techniques are being increasingly employed for text processing tasks, the need for large training datasets has become a major bottleneck for their application. Manual generation of large scale datasets tailored to each task is a time consuming and expensive process, which necessitates the need for automated generation of training datasets. In this work, we turn our attention towards creation of training datasets for named entity recognition (NER) in the context of cultural heritage domain. NER plays an important role in many natural language processing systems. Most NER systems are typically limited to a few common named entity types, such as person, location, and organization. However, for cultural heritage resources, such as digitized art archives, the recognition of fine-grained entity types such as titles of artworks is of high importance. Current state of the art tools are unable to adequately identify artwork titles due to unavailability of relevant training datasets. We analyse the particular difficulties presented by this domain and motivate the need for quality annotations to train machine learning models for identification of artwork titles. We present a framework with heuristic based approach to create high-quality training data by leveraging existing cultural heritage resources from knowledge bases such as Wikidata. Experimental evaluation shows significant improvement over baseline for NER performance for artwork titles when models are trained on the dataset generated using our framework.

Keywords: cultural heritage collections, training data generation, named entity recognition, weakly-supervised learning

1. Introduction

Deep learning models have become popular for natural language processing (NLP) tasks in recent years. This is accounted to the superior performance achieved by the neural networks based techniques on a wide range of NLP problems as compared to the traditional statistical techniques. State-of-the-art results have been achieved by deep learning approaches for named entity recognition, question answering, machine translation and sentiment analysis — among others. As supervised model training methods have become ubiquitous, the availability of training data has emerged as one of the major challenges for their success. For standard NLP tasks, the research community has been leveraging a set of common and widely distributed training datasets that are tailored to the respective tasks [1–4]. However, such training datasets are not generically applicable to variations of the standard problems or to different domains. Without relevant good quality training data, even the most successful and innovative deep learning architectures cannot hope to achieve good results.

In this work, we focus on the named entity recognition (NER) task which seeks to identify the boundaries of text that refer to named entities and to categorize the found named entities into different types. NER serves as an important step for various semantic tasks, such as knowledge base creation [5], machine translation [6], relation extraction [7] and question answering [8], etc. Most NER efforts are restricted to only on a few common categories of named entities, i.e., person, organization, location, and date. This is generally referred to as coarse-grained NER, as compared to the fine-grained NER or FiNER which aims to classify the entities into several more entity types [9, 10].
FiNER helps to determine precisely the semantics of the identified entities and this is desirable for many downstream tasks. Previous research has demonstrated that the performance of the relation extraction task, that takes the named entities as input, is boosted by a considerable margin when supplied with a larger set of FiNER types as opposed to the four types [9, 11]. Question answering systems have also been shown to benefit from fine-grained entity recognition as it helps to narrow down the results based on expected answer types [12, 13]. Fine-grained NER is also essential for domain-specific NER, where various different named entity categories are of higher importance and relevance depending on the domain itself. E.g., for a company dealing with financial data, named entity types such as Bank, Loans, etc. would be important to detect and classify, while for biomedical data, the names of Proteins, Genes etc would be important to correctly identify.

Most of the recent neural network based NER models have been trained on a few well-established corpora available for the task such as the CoNLL datasets [2, 14] or OntoNotes [15]. Although these systems attain state-of-the-art results for the generic NER task, their performance and utility for identifying fine-grained entities is essentially limited due to the specific training of the models. Thus, it comes as no surprise that it has been a challenge to adapt NER systems for identifying fine-grained and domain-specific named entities with reasonable accuracy [16, 17]. This is especially true for cultural heritage data where the artefacts serve as one of the most important named entity categories. Recently, there has been a surge in the availability of digitized cultural data with the principles of linked open data gaining momentum in the cultural heritage domain [18]. Initiatives such as OpenGLAM and flagship digital library projects such as Europeana aim to enrich open knowledge graphs with cultural heritage data by improving the coverage of the topics related to the cultural domain. Efforts have been made to digitize historical archives in various domains. This is especially true for the art domain where a large collection of raw texts are yet to be explored/analysed. These collections consist of art related texts such as auction catalogues, art books and exhibition catalogues [19, 20]. In such resources, cultural objects, mainly artworks, are often described with help of unstructured text narratives. The identification and extraction of the mentions of artworks from such text descriptions facilitates search and browsing in digital resources, helps art historians to track the provenance of artworks and enables wider semantic text exploration for digital cultural resources.

While several previous works on FiNER have defined entity types ranging from hundreds [9, 10, 21] to thousands [22] of different types, they are not specifically catered to the art domain. Ling et al. [9] have defined 112 named entity types from generic areas. Similarly to Gillick et al. [10], they added finer categories for certain types such as actor, writer, painter or coach that are sub-types of the Person class, and city, country, province, island, etc. that belong to the Location type. They also added other new entity types such as Building and Product that have their own sub-types. Although these works have defined certain entity types that are domain-specific, such as disease, symptom, drug for the biomedical domain and music, play, film, etc. for the art domain, an exhaustive list of all important entity types for different domains is not achievable in a generic fine-grained NER pipeline. As per the authors’ knowledge, none of the existing efforts have explicitly considered and added an artwork such as painting or sculpture as a named entity type to their type list. As such, there is no available large scale annotated data for training supervised machine learning models to identify artwork titles as named entities.

The focus of this work is to propose techniques for generating large, good quality annotated datasets for training FiNER models. We investigate in detail the identification of mentions of artworks, as a specific type of named entity, from digitized art archives. To this end, we leverage existing art resources that are integrated in popular knowledge bases, such as Wikidata [23]. Further, we augment the training data with silver standard annotations derived from well-structured and clean texts from Wikipedia articles referring to artworks. These silver standard annotations provide important textual features and patterns that are indicative of artwork titles in free form texts. Our evaluation demonstrates substantial improvement in NER performance (doubling the F1 score) when trained with the high-quality annotations generated through our methods. This confirms the effectiveness of our methods while also validating our approach to focus on generating high-quality training data that is essential for domain-specific tasks.
We discuss the specific challenges of identifying artwork titles in the next section and motivate the necessity of generation of training data for this problem.

2. Named Entity Recognition for Artworks

Identification of mentions of artworks seems, at first glance, to be no more difficult than detecting mentions of persons or locations. But the special characteristics of these mentions makes this a complicated task which requires significant domain expertise to tackle. We introduce the named entity type artwork. Artworks are typically referred to by their titles, these titles could have been assigned by artists or, in the case of certain old and ambiguous artworks, by collectors, art historians, or other domain experts. Due to the ambiguities that are inherent in artwork titles, their identification from texts is a challenging task. As an example, consider the painting titled ‘Girl before a mirror’ by famous artist Pablo Picasso. This title merely describes in an abstract manner what is being depicted in the painting and thus, it is hard to identify it as a named entity without knowing the context of its mention. Similarly, consider the painting with the title ‘Head of a woman’ — such phrases can be hard to be distinguished as named entities from the surrounding text due to their generality. Yet, such descriptive titles are common in the art domain, as are abstract titles such as ‘untitled’.

To circumvent ambiguities present in art-related documents for human readers, artwork titles are typically formatted in special ways — they are distinctly highlighted with capitalization, quotes, italics or boldface fonts, etc. which provide the required contextual hints to identify them as titles. However, the presence of these formatting cues cannot be assumed or guaranteed, especially in texts from art historical archives, due to adverse effects of scanning errors on the quality of digitized resources [24]. Moreover, the formatting cues for artwork titles might vary from one text collection to the other. Therefore, the techniques for identifying the titles in digitized resources need to be independent of formatting and structural hints, making the task even more complex. Moreover, the quality of digitized versions of historical archives is adversely affected by the OCR scanning limitations and the resulting data suffers from spelling mistakes as well as formatting errors. The issue of noisy data further exacerbates the challenges for automated text analysis, including the NER task [25]. For this work, the underlying dataset is a large collection of art historical documents that have been recently digitized. For reference, a sample document4 from a similar collection is shown in Fig. 1a. The collection consists of different types of documents: auction catalogues, full texts of art books related to particular artists or art genres, catalogues of art exhibitions and other documents. The auction and exhibition catalogues contain semi-structured and unstructured texts that describe artworks on display, mainly paintings and sculptures. Art books may contain more unstructured text about the origins of artworks and their creators. Fig. 1b shows the proportion of the different kinds of documents in the dataset. The pages of these catalogues and books were scanned with OCR and each page was converted to an entry stored within an elastic search index. Due to the limitations of OCR, the dataset did not retain its rich original formatting information which would have been very useful for analysis. In fact, the data suffers from many spelling and formatting mistakes that need to be appropriately handled.

We want to illustrate the difficulties that arise when trying to recognize artwork mentions in practice. There are three types of errors that can be distinguished — Failure of detection of a artwork named entity, incorrect detection of the named entity boundaries, and incorrect tagging of the artwork with a wrong type.

2.1. Incorrectly Missed Artwork Title

Many artwork titles contain generic words that can be found in dictionary. This poses difficulties in the recognition of titles as named entities. E.g., a painting titled ‘A pair of shoes’ by Van Gogh can be easily missed while searching for named entities in unstructured text. Such titles can only be identified if they are appropriately capitalized or highlighted, however this cannot be guaranteed for all languages and in noisy texts.

2.2. Incorrect Artwork Title Boundary Detection

Often, artworks have long and descriptive titles, e.g., a painting by Van Gogh titled ‘Head of a peasant woman with dark cap’. If this title is mentioned in text without any formatting indicators, it is likely that the

---

boundaries may be wrongly identified and the named entity be tagged as ‘Head of a peasant woman’, which is also the title of a different painting by Van Gogh. In fact, Van Gogh had created several paintings with this title in different years. For such titles, it is common that location or time indicators are appended to the titles (by the collectors or curators of museums) in order to differentiate the artworks. However, such indicators are not a part of the original title and should not be included within the scope of the named entity.

On the other hand, for the painting titled ‘Black Circle (1924)’ the phrase ‘(1924)’ is indeed a part of the original title and should be tagged as such. There are many other ambiguities for artwork titles, particularly for older works that are typically present in art historical archives.

2.3. Incorrect Type Tagging of Artwork Title

Even when the boundaries of the artwork titles are identified correctly, they might be tagged as the wrong entity type. This is especially true for the artworks that are directly named after the person whom they depict. The most well-known example is that of ‘Mona Lisa’, which refers to the person as well as the painting by Da Vinci that depicts her. There are many other examples such as Picasso’s ‘Jaqueline’, which is a portrait of his wife Jaqueline Rogue. Numerous old paintings are portraits of the prominent personalities of those times and are named after them such as ‘King George III’, ‘King Philip II of Spain’, ‘Queen Anne’ and so on. Many painters and artists also have their self-portraits named after them — such artwork titles are likely to be wrongly tagged as the person type in the absence of contextual clues. Apart from names of persons, paintings may also be named after locations such as ‘Paris’, ‘New York’, ‘Grand Canal, Venice’ and so on and may be incorrectly tagged as location. Yet another type of ambiguity involving both incorrect boundaries and wrong tagging can occur when paintings with long titles contain phrases that match with other named entities, consider the title ‘Lambeth Palace seen through an arch of Westminster Bridge’ which is an artwork by English painter Daniel Turner. In this title, ‘Lambeth Palace’ and ‘Westminster Bridge’ are both separately identified as named entities of type location, however, the title as a whole is not tagged as any named entity at all.

The above examples demonstrate the practical difficulties for automatic identification of artwork titles. In our dataset, we encountered many additional errors due to noisy text of scanned art historical archives.

Due to the innate complexity of this task, NER models need to be trained with domain-specific named entity annotations, such that the models can learn important textual features to achieve the desired results. As such, the unavailability of high-quality training data for the cultural heritage domain is one of the biggest hindrances for this task. We address this gap by proposing techniques for generating annotations for NER via a semi-automated approach from a large corpus of art related documents.

Central to the idea of identification of the mentions of artworks is the task of mapping different mentions of the same artwork or disambiguation of distinct artworks having the same name to their correct artwork. Although an interesting challenge, this work does not address the issue of named entity linking, where the identified artworks would be mapped to the corresponding instance on existing knowledge graph and we leave this as future work.
In the next section, we compare and contrast the research efforts related to our work before describing our approach to tackle these challenges in Section 4.

3. Related Work

We discuss the related work under different categories, we start with a general overview of previous work on NER and the need for annotated datasets, followed by a discussion on domain specific and fine-grained NER in the context of cultural heritage resources. Then we present the related efforts for automated training data generation for machine learning models, particularly for NER.

NER being important for many NLP tasks, has been the subject of numerous research efforts. Several prominent systems have been developed that have achieved near human performance for the few most common entity types on certain datasets. Previously, the best performing NER systems were trained through feature-engineered techniques such as Hidden Markov Models (HMM), Support Vector Machines (SVM) and Conditional Random Fields (CRF) [26–29]. In the past decade, such systems have been succeeded by neural network based architectures that do not rely on handcrafted features to identify named entities correctly. Many architectures leveraging Recurrent Neural Networks (RNN) for word level representation [30–32], and Convolutional Neural Networks (CNN) for character level representation [33–35] have been proposed recently. The latest neural-networks-based NER models use a combination of character and word level representations along with variations of features from previous approaches. These models have achieved state of the art results on multilingual CoNLL 2002 and 2003 datasets [36, 37].

However, all these systems are dependent on a few prevalent benchmark datasets that provide gold standard annotations for training purposes. These benchmark datasets were manually annotated using proper guidelines and domain expertise. E.g., the CoNLL and OntoNotes datasets, that were created on news-wire articles, are widely shared among the research community. Since these NER systems are trained on a corpus of news articles they perform well only for comparable datasets. In most cases, these systems fail to adapt well to new domains and different named entity categories [16, 17].

Domain specific NER. There is prior work for domain specific NER, such as for the biomedical domain. NER systems have been used to identify the names of drugs, proteins and genes [38–40]. But since these techniques rely on specific resources such as carefully curated lists for drug names [41] or biology and microbiology NER datasets [42, 43], they are highly specific solutions geared towards biomedical domain and cannot be applied directly to cultural heritage data.

In the absence of gold standard NER annotation datasets, the adaptation of existing solutions to the art and cultural heritage domain faces many challenges, some of them being unique to this domain. Seth et al. [44] discuss some of these difficulties and compare the performance of several NER tools on descriptions of objects from the Smithsonian Cooper-Hewitt National Design Museum in New York. Segers et al. [45] also offer an interesting evaluation of the extraction of event types, actors, locations, and dates from unstructured text present in the management database of the Rijksmuseum in Amsterdam. However, their test data contains Wikipedia articles which are well-structured and more suitable for extraction of named entities. On similar lines, Rodríguez et al. [25] discuss the performance of several available NER services on a corpus of mid-20th-century typewritten documents and compare their performance against manually annotated test data having named entities of types people, locations, and organizations. Ehrmann et al. [46] offer a diachronic evaluation of various NER tools for digitized archives of Swiss newspapers. However, none of the existing works have focused on the task of identifying artwork titles which are one of the most important named entities for the art domain. Moreover, previous works have merely compared the performance of existing NER systems, whereas in this work, we aim to improve the performance of NER systems for cultural heritage by the generating domain-specific high-quality training data. Recently, there has been increasing effort to publish cultural heritage collections as linked data [20, 47, 48], however, to the best of our knowledge, there is no annotated dataset for NER available for this domain which is the focus of this work.

Training Data Generation. For the majority of the previous work related to NER, the primary research focus has been on the improvement of the model architectures with the help of novel machine learning and neural networks based approaches. The training as well as evaluations for these models are performed
on the publicly available popular benchmark datasets. This approach is not feasible for targeted tasks, such as for the identification of artwork titles due to the requirement of specialized model training on related datasets. Manual curation of gold standard annotations for large domain-specific corpus is expensive in terms of human labour and cost, while also requiring significant domain expertise. Hence our work complements the efforts of NER model improvements by focusing on the automated generation of training datasets for these models.

In [49], the authors attempt to aid the creation of labeled training data in weakly-supervised fashion by a heuristic based approach. Snorkel [50] is an open source system that enables training data creation by the user manually specifying a set of heuristics, labelling functions and patterns. While this system generates good quality annotations, it is task and domain specific, where different rules or functions designed by experts are required for every task, whereas our framework is generic enough to be applicable to other domains without the need of domain expertise. Other works that depend on heuristic patterns along with user input are [51, 52]. Similar to our approach, Mints et al. [7] leveraged Freebase knowledge base and used distant supervision for training relation extractors.

In the context of generating training datasets for NER, previous works have exploited the linked structure of Wikipedia to identify and tag the entities with their type, thus creating annotations via distance supervision [53, 54]. Ghaddar and Langlais further extended this work by adding more annotations from Wikipedia in [55] and adding fine-grained types for the entities in [56]. However, these techniques are only useful in a very limited way for the cultural heritage domain, since Wikipedia texts do not contain sufficient entity types relevant to this domain. Instead of using a generic and cleanly formatted text like Wikipedia to annotate many different entity types, as is done for fine-grained NER in previous works, our focus is to instead annotate a domain specific corpus for relevant entities. Our approach is able to work with noisy data from digitized art archives to automatically create annotations for artwork titles. We propose a framework to generate high-quality training corpus in a scalable and automated manner and demonstrate that NER models can be trained to identify mentions of artworks with notable performance gains.

4. Annotating Complex Named Entity Types

In this section we discuss our three stage framework for generating high-quality training data for the NER task without the need for manual annotations (Fig. 2). These techniques were geared towards tackling the challenges presented by noisy corpora that are typical of art historical archives, although they can be applicable for other domains as well. The framework can take structured or unstructured data as input and progressively add annotations for artwork named entity at each stage. Though our input dataset is multilingual, we put focus on the English texts for the sake of simplicity in this work. We describe the three stages of the framework and the output dataset at each stage.

4.1. Stage I - Tagging Artworks as Named Entities

In order to match and correctly tag the artworks present in our corpus as named entities, we leveraged cultural resources that have been integrated into the public knowledge bases. In the first stage, we collected available resources from Wikidata to generate a large entity dictionary or gazetteer of artwork titles in an automatic way. Integrating other sources, such as art-related ontologies or lists from museum resources is also possible. To generate the entity dictionary for titles, Wikidata was queried with the Wikidata Query Service for names of artworks, specifically for names of paintings and sculptures. Since our input dataset was inherently multilingual, there were many instances where the original non-English titles of paintings were mentioned in the texts. In order to match such titles, we added all the alternate names of the paintings and sculptures to our list belonging to the 7 major languages present in the dataset apart from English (French, German, Italian, Dutch, Spanish, Swedish and Danish). A large variety of artwork titles were obtained from Wikidata, with the shortest title belonging to a painting being just a few characters ("C-B-1"), while the longest title having 221 characters in total ("Predella Panel Representing the Legend of St. Stephen... ").

Quite a few of the titles were highly generic, for instance, 'Italian', 'Winter', 'Landscape' etc., therefore, we filtered out the titles having only one word from the list. Since several artwork titles are identical to location names which can lead to errors while tagging

5https://query.wikidata.org/
the named entity to the correct type, such titles were also ignored. The large variety and ambiguity observed in the titles extracted from Wikidata further confirmed that the NER for artwork titles is a non-trivial task. A combined list of approximately 15,000 titles in different languages were obtained, majority of them being in English. Due to inconsistencies in the capitalization of the words in the title found on Wikidata, as well as in the mention of titles in our dataset, the titles had to be uniformly lower-cased to enable matching. The simple technique of matching the dictionary items over the words in our dataset to tag them as artwork entities did not yield reasonable results. This was mainly due to the generality of the titles. As an example, consider the painting title ‘three girls’. If this phrase would be searched over the entire corpus, there could be many incorrect matches where the text would perhaps be used to describe some artwork instead of referring to the actual title. To circumvent this issue of false positives, we first extracted named entities of all categories as identified by a generic NER model (details in section 4.4). Thereafter, those extracted named entities that were successfully matched with an artwork title in the entity dictionary, were considered as artworks and their category was explicitly tagged as artwork. Even though some named entities were inadvertently missed with this approach, it facilitated the generation of high-precision annotations from the underlying dataset from which the NER model could learn useful features.

4.2. Stage II - Improving Named Entity Boundaries

As discussed in Section 2.2, there can be many ambiguities due to partial matching of artwork titles. Due to the limitations of the naive NER model, there were many instances where only a part of the full title of artwork was recognized as a named entity from the text, thus it was not tagged correctly as such. To improve the recall of the annotations, we attempted to identify the partial matches and extend the boundaries of the named entities to obtain the complete and correct titles in the second stage.

For a given text, a separate list of matches with the artwork titles in entity dictionary over the entire text was maintained as spans (starting and ending character offsets), in addition to the extracted named entities. It is to be noted that the list of spans included many false positives due to matching of generic words and phrases that were not named entities. The overlaps between the two lists were considered, if a span was a super-set of a named entity, the boundary of the identified named entity was extended as per the span offsets. For example, from the text ‘..The subject of the former (inv. 3297) is not Christ before Caiaphas, as stated by Birke and Kertesz, but Christ before Annas..’ , the named entities ‘Christ’, ‘Caiaphas’ and ‘Annas’ were separately identified initially. However, they were correctly updated to ‘Christ before Caiaphas’ and ‘Christ before Annas’ as artwork entities after the boundary corrections. Through this technique, many missed mentions of artwork titles were added to the training dataset, thus improving the recall of the annotations and the overall quality of the dataset.

4.3. Enhancements with Silver Standard Train Data

Despite efforts for high precision in stage I, one of the major limitations of generating named entity annotations from art historical archives is the presence of errors in the training data. Since the input dataset consists of noisy text, it is inevitable that there would be errors in the matching of artwork titles as well as in the recognition of the entity boundaries. To enable an NER model to learn the textual indicators present in the dataset for identification of artworks, this stage further augmented our training dataset with clean and well-structured silver standard\textsuperscript{6} annotations derived from Wikipedia articles that proved very useful for NER training. To find such sentences, firstly, we searched for the Wikipedia pages of all the artwork titles in English wherever applicable; a total of 2808 pages were found. We then extracted the relevant sentences that mentioned the artwork title from these pages. To obtain more sentences, we also leveraged the link struc-

\textsuperscript{6}The examples are not manually annotated by experts but the annotations are derived in an automatic fashion, therefore silver standard data is often lower in quality compared to gold standard data.
ture of Wikipedia and mined relevant sentences from the different Wikipedia articles that, in turn, referred to a Wikipedia article of an artwork. Several previous works have utilized the anchor texts and the tagged categories present in Wikipedia articles to transform sentences into named entity annotations [57–59]. We followed a somewhat similar approach — for each Wikipedia page referring an artwork, the back-links, i.e. the URLs of the pages that referred to this page were collected. The pages were searched for the relevant sentences that contained an outgoing link to the Wikipedia page of the artwork, while also making sure that anchor text of the outgoing link was identical to the title of the artwork. These sentences were extracted and the anchor texts of the sentences was tagged as an artwork, serving as accurate annotations for this category. In this stage, a total of 1628 sentences were added as silver standard annotation data to the training set. The process is illustrated in Fig. 3. This data provided correct and precise textual patterns that were highly indicative of the artwork titles and led to a further boost in training data quality.

4.4. Baseline NER Model

None of the existing NER systems can identify titles of artworks as named entities out of the box. The closest NER category to artwork titles was found in the SpaCy7 library as work_of_art. This category refers not only to artworks such as paintings and sculptures, but also covers a large variety of cultural heritage objects including movies, plays, books, songs etc. For the lack of alternatives, we have leveraged this NER category in our work for setting up a naive baseline with which we compare the improvements in NER performance.

The SpaCy library for natural language processing was employed for tokenization and chunking of the texts before the identification of the named entities. The pre-trained English model of SpaCy has been trained on Ontonotes5 dataset8 which consists of different types of texts including telephone conversations, news-wire, newsgroups, broadcast news etc. Since this dataset is considerably different from historical art document collections, the pre-trained NER model showed poor performance for named entity recognition in the cultural heritage domain, even for the common named entity types (person, location and organization). With regards to artwork titles, very few were identified as named entities and many among those were wrongly tagged as names of persons or locations, instead of being correctly categorized as work_of_art. The pre-trained SpaCy NER model will be referred to as the baseline model. In order to improve the identification of artwork entities, training on high-quality annotated training datasets is imperative and for this purpose, the baseline NER model was leveraged for re-training. Due to the steep costs and efforts of human annotations, we aimed to generate a large corpus of annotated data in a semi-automated fashion from our dataset. It is to be noted that the techniques for improving the quality of NER training data

7SpaCy: https://spacy.io/, version 2.1.3
8https://catalog.ldc.upenn.edu/LDC2013T19
that are proposed in this work are independent of the NER model used for the evaluation. Thus, SpaCy is merely a tool which can be substituted with any other re-trainable NER system.

5. Evaluations and Discussion

In this section, we discuss the details of our experimental setup and present the performance results of the NER models when trained on annotated dataset generated with our approach.

5.1. Experimental Setup

The input dataset to our framework consisted of art related texts in many different languages including English, French, German, Italian, Dutch, Spanish, Swedish and Danish among others. After filtering out English texts and performing initial preprocessing, including the removal of erroneous characters, the dataset consisted of 19310429 sentences. This included partial sentences such as artwork size entries as well as well-formed sentences describing the artworks. This sizeable, noisy input dataset was transformed into annotated NER data through the three stages of our framework as described in Section 4.

In order to evaluate and compare the impact on NER performance with improvements in quality of the training data, we trained the baseline NER model for the new entity type artwork on different variants of training data. Apart from the Baseline NER model (pre-trained with no ‘artwork’ annotations), the following versions of the training data having ‘artwork’ annotations were used for the model training -

**WD** — Training dataset with annotations obtained with matching of Wikidata titles, along with named entity boundary corrections (Stage I and II)

**WP** — Training dataset generated only with silver standard annotations derived from Wikipedia in Stage III

**WD-WP** — Training dataset with augmentation of the above two datasets, i.e. combined output of Stage I, II and III of the framework.

The statistics of the these datasets are shown in Table 1. (In WP dataset, each sentence had corresponding annotation). An NER model was trained on each of the above datasets for 10 epochs, with the training data batched and shuffled before every iteration. The performance of the trained NER models was compared with the Baseline NER model (which was pre-trained without any specific annotations for artwork titles). Since the named entity type artwork was not applicable for the baseline model, a match with the entity type work_of_art was considered as a true positive. In the absence of a gold standard dataset for NER for artwork titles, we performed manual annotations and generated a test dataset on which the models could be suitably evaluated.

5.2. Manual Annotations for Test Dataset

For generating a test dataset, a set of texts were chosen at random from the dataset, while making sure that this text was representative of the different types of document collections in the overall corpus. This test data consisted of 544 entries (with one or more sentences per entry) and was carefully excluded from the training dataset. The titles of paintings and sculptures mentioned in this data were then manually identified and tagged as named entities of artwork type. The annotations were performed by two non-expert annotators independently of each other in 3 – 4 person hours with the help of Enno tool and their respective annotations were compared afterwards. The task of manual annotation was found challenging due to the inherent ambiguities in the dataset (Section 2) and lack of domain expertise. The annotators disagreed on the tagging of certain phrases as titles on multiple occasions. For example, in the text snippet “An earlier, independent watercolor of almost the same view can be dated to circa 1830 (Stadt Bernkastel-Kues; see C. Powell, Turner in Germany, exhibition catalogue, London, Tate Gallery, 1995-96, pp. 108-9, no. 23> illustrated in color).”, the artwork mention ‘Stadt Bernkastel-Kues’ was missed by one of the annotators. The correct boundaries of the artworks was also disagreed in some cases, such as in the text “Claude Monet, Rouen Cathedral, Facade, 1894, Oil on canvas [W.1356]. Museum of Fine Arts, Boston” - the artwork title could be ‘Rouen Cathedral, Facade’ or ‘Rouen Cathedral’. The inter-annotator agreement in

<table>
<thead>
<tr>
<th>Table 1: Statistics of Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>Sentences</td>
</tr>
<tr>
<td>Annotations</td>
</tr>
<tr>
<td>Unique entities</td>
</tr>
</tbody>
</table>

https://github.com/HPI-Information-Systems/enno
10

N. Jain et al. / Overcoming Shortage of Training Data for Cultural Heritage Analysis

2

3
terms of the Fleis-kappa and Krippendorf-kappa scores
were calculated to be -1.86 and 0.61 respectively. (A
negative Fleis-kappa score indicates poor agreement,
while Krippendorf-kappa values for data should be
above 0.667 to be considered useful.) The poor inter-
annotator agreement reflected by these scores reaff-
irmed that the task of annotating the artwork titles is
difficult, even for humans. In order to obtain the gold
standard test dataset for the evaluation of NER mod-
els, the disagreements were manually sorted out with
the help of web search and a total of 144 entities were
positively tagged as artwork.

5.3. Evaluation Metrics

The performance of NER systems is generally mea-
sured in terms of precision, recall and F1 scores. The
correct matching of a named entity involves the match-
ing of the boundaries of the entity (in terms of charac-
ter offsets in text) as well as the tagging of the named
title to the correct category. The strict F1 scores
for NER evaluation were used in the CoNNL 2003
shared task10, where the entities’ boundaries were
matched exactly. The MUC NER task11 allowed for re-
laxed evaluation based on the matching of left or right
boundary of an identified named entity. In this work,
the evaluation of NER was performed only for artwork
titles and therefore, it was sufficient to check only
for the boundary matches of the identified entities.
Since there are many ambiguities involved with entity
boundaries of artwork titles, as discussed in Section
2.2, we evaluated the NER models with both strict met-
rics based on exact boundary match, as well as the re-
laxed metrics based on partial boundary matches. The
relaxed F1 metric allowed for comparison of the enti-
ties despite errors due to wrong chunking of the named
titles in the text. Precision, recall, F1 as well as accu-

4

5
ccuracy scores obtained for the NER models trained with
different training dataset variants are shown in Table 2.

5.4. Results and Discussion

The results demonstrated definitive improvement in
performance for the NER models that were trained
with annotated data as compared to the baseline perfor-
mance. Since the relaxed metrics allowed for flexible
matching of the boundaries of the identified titles, they

\[\text{Table 2}
\]

<table>
<thead>
<tr>
<th>Train Dataset Stage</th>
<th>Strict</th>
<th>Relaxed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline –</td>
<td>.14 .06</td>
<td>.24 .22 .08 .37</td>
</tr>
<tr>
<td>WD I, II</td>
<td>.23 .22</td>
<td>.61 .39 .41 .40 .68</td>
</tr>
<tr>
<td>WP III</td>
<td>.17 .13</td>
<td>.54 .38 .30 .34 .62</td>
</tr>
<tr>
<td>WD-WP I, II, III</td>
<td>.39 .36</td>
<td>.71 .56 .53 .55</td>
</tr>
</tbody>
</table>

Fig. 4. NER Performance with Different Training Data Sizes

were consistently better than the strict matching scores
for all cases. The training data obtained from Stage
I and II i.e. the WD dataset enabled an improvement
in NER performance due to the benefit of domain-
specific and entity-specific annotations generated from
the Wikidata entity dictionaries, along with the boost
from additional annotations by the correction of entity
boundaries. To gauge the benefits from the silver stan-
dard annotations from Wikipedia sentences, a model
was trained only on these sentences(WP). It can be
seen that the accuracy measure of this model was quite
high despite the small size of the dataset, indicating
the positive impact of the quality of the annotations.
The model trained on the final training dataset obtained
through our framework, consisting of all the annota-
tions obtained from the three stages(WD-WP), showed
the best overall performance with significant improve-
ment across all metrics. This demonstrated the impor-
Table 3
Analysis of Model Annotations after Training

<table>
<thead>
<tr>
<th>#</th>
<th>Text</th>
<th>Identified Title</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>'Figure 39. On the Terrace. Panel, 17.7 x 18 cm. The Cleveland Museum of Art, Bequest of Clara Louise Gehring Bickford, 1986.68. Photo: Courtesy of the Museum.'</td>
<td>On the Terrace</td>
<td>True Positive</td>
</tr>
<tr>
<td>2.</td>
<td>'...as in End of a Gambling Quarrel (Fig. 45), where the furniture is overturned, one chair projecting in the very picture surface and the cards are strewn.'</td>
<td>End of a Gambling Quarrel</td>
<td>True Positive</td>
</tr>
<tr>
<td>3.</td>
<td>'he owned a painting entitled The Little Nephew of Rameau (1858), a rare instance of Meissonier making a literary allusion.'</td>
<td>The Little Nephew of Rameau</td>
<td>True Positive</td>
</tr>
<tr>
<td>4.</td>
<td>'Figure 34. The Inn Door in the Saint-Germain Forest. Panel, 17 x 23 cm. Paris, Musee d'Orsay.'</td>
<td>The Inn Door</td>
<td>Partial Match</td>
</tr>
<tr>
<td>5.</td>
<td>'Among the other works in Davis's private collection was The Grand Canal with Ca' Pesaro by Francesco Guardi, sold at Christie's, London.'</td>
<td>The Grand Canal with Ca' Pesaro</td>
<td>Partial Match</td>
</tr>
<tr>
<td>6.</td>
<td>'the writings of contemporaries like Alexandre Dumas, whose The Three Musketeers was published as a novel in 1844 and performed as a play in 1845,'</td>
<td>The Three Musketeers</td>
<td>False Positive</td>
</tr>
<tr>
<td>7.</td>
<td>'Property from the Collection of William And Eleanor Wood Prince, CHICAGO, ILLINOIS'</td>
<td>The Collection of Wood Prince</td>
<td>False Positive</td>
</tr>
<tr>
<td>8.</td>
<td>'..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 Maast and Castel (no. 904).''</td>
<td>William And Eleanor Wood Prince</td>
<td>False Negative</td>
</tr>
<tr>
<td>9.</td>
<td>'like the crumpled paper and feather broken from a pen in Young Man Working or the green leaf fallen from the fruit plate in The Confidence.'</td>
<td>The Confidence</td>
<td>False Negative</td>
</tr>
</tbody>
</table>

Of training on high-quality annotation datasets for named entity recognition. Our approach to generate such annotations in a semi-automated manner from a domain-specific corpus is an important contribution towards this direction. Moreover, the remarkable improvement for NER performance achieved for a novel and challenging named entity of type artwork, proves the effectiveness of our approach.

5.5. Impact of Training Data Size

To inspect the effect of the size of the generated training data on NER performance, we varied the dataset size and performed the model training on progressively increasing sizes of training data. We randomly sampled smaller datasets from the overall training dataset in the range 10 per cent to 90 per cent and plotted the performance scores of the trained models (averaged over 10 iterations) as shown in Fig. 4a and 4b. It can be seen that all the scores show a general upward trend as the training data size increases. The best scores were achieved with the entire training dataset that was obtained as output from the framework.

5.6. Error Analysis

A closer inspection of the performance of NER models revealed interesting insights. Some example annotations performed by the trained NER model are shown in Table 3. As discussed in Section 2, it is intrinsically hard to identify mentions of artworks from the digitized art archives. The noise present in the text further exacerbates the problem. In the supervised learning setting, a neural network model is expected to learn patterns based on the annotations that are fed to it during the training phase. Based on this fact, the third stage of our framework incorporates the silver standard sentences from Wikipedia so as to provide clean and precise artwork annotations. From such annotations, the model could learn the textual patterns that are indicative of the mention of an artwork title. An evaluation of the annotations performed by model on our test dataset shows that the model was indeed able to learn such patterns. For example, in Text 1 from an exhibition catalogue, the model was able to identify the title 'On the Terrace' correctly. Similarly, from the Text 2, the title 'End of a Gambling Quarrel' was identified. It can be seen from these examples that the model is able to understand cues such as the presence of 'Figure' or 'Fig.' in the vicinity of the title. Not only this, the model is able to understand that textual patterns such as '...a painting entitled...' are usually followed by the title of the artwork, as shown in Text 3.

Even after performing the checking of the entity boundaries during the generation of annotation dataset, the model still made errors in entity recognition in terms of marking the boundaries. This is illustrated by Text 4 and 5 in Table 3. Given the particular use case of noisy art collections and the ambiguities inherent in artwork titles, this is indeed a hard problem to tackle. The relaxed metrics consider partial matches as positive matches and favour the trained NER model in such cases.
There were also a few instances where the model
wrongly identified a named entity of a different type
as artwork. This is likely to happen when the entity is
of a similar type, such as the title of a book or a play,
such as in the Text 6. In some cases, the names of per-
sons is misleading to the model and wrongly tagged as
artwork, such as in Text 7. Finally, Texts 8 and 9 show
some examples where the model simply could not de-
tect the titles of artworks due to lack of hints or fami-
lar patterns to rely upon. In spite of the difficulties for
this specific entity type, it is encouraging to note the
improvement of performance of the NER model, mak-
ing the case for the usefulness of the generated training
data by our framework.

6. Conclusion

In this work we proposed a framework to gener-
ate a large number of annotations for identifying art-
work mentions from art collections. We motivated the
need for NER training on high-quality annotations and
proposed techniques for generating the relevant train-
ing data for this task in semi-automated manner. Ex-
perimental evaluations showed that the NER perfor-
ance can be significantly improved by training on
high-quality training data generated with our methods.
This indicates that even for noisy datasets, such as
digitized art archives, supervised NER models can be
trained to perform well. Furthermore, our approach is
not limited to the cultural heritage domain but can be
adapted for other domains, where there is also shortage
of annotated training data. As future work we would
like to apply our techniques for named entity recog-
nition to other important entities and perform entity-
centric text exploration for cultural heritage resources.
It would be interesting to leverage named entities to
mine interesting patterns about artworks and artists,
which may facilitate the creation of a comprehensive
knowledge base for this domain.

References

positionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language pro-
[4] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng and C. Potts, Learning word vectors for sentiment analysis, in: Pro-
ceedings of the 49th annual meeting of the association for com-
putational linguistics: Human language technologies-volume 1, Association for Computational Linguistics, 2011, pp. 142–
150.
phy, T. Strohmann, S. Sun and W. Zhang, Knowledge vault: A web-scale approach to probabilistic knowledge fusion, in:
Proceedings of the 20th ACM SIGKDD international confer-
one on Knowledge discovery and data mining, ACM, 2014, pp. 601–610.
N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens et al., Moses: Open source toolkit for statistical machine translation,
in: Proceedings of the 45th annual meeting of the association for com-
[7] M. Mintz, S. Bills, R. Snow and D. Jurafsky, Distant supervi-
sion for relation extraction without labeled data, in: Proceed-
ings of the Joint Conference of the 47th Annual Meeting of
the ACL and the 4th International Joint Conference on Natu-
ral Language Processing of the AFNLP: Volume 2-Volume 2,
Association for Computational Linguistics, 2009, pp. 1003–
1011.
[8] T. Lin, O. Etzioni et al., No noun phrase left behind: detecting
and typing unlinkable entities, in: Proceedings of the 2012 joint
cconference on empirical methods in natural language process-
ing and computational natural language learning, Association
[9] X. Ling and D.S. Weld, Fine-grained entity recognition,
in: Twenty-Sixth AAAI Conference on Artificial Intelligence,
2012.
[10] D. Gillick, N. Lazie, K. Ganchev, J. Kirchner and D. Huynh,
distantly supervised relation extraction with linked arguments,
in Natural Language Processing (EMNLP), 2014, pp. 1891–
1901.
[12] L. Dong, F. Wei, H. Sun, M. Zhou and K. Xu, A hybrid
neural model for type classification of entity mentions, in:
Twenty-Fourth International Joint Conference on Artificial In-
telligence, 2015.
[13] X. Li and D. Roth, Learning question classifiers, in: Proceed-
ings of the 19th international conference on Computational
linguistics-volume 1, Association for Computational Linguis-
Shared Task: Language-Independent Named Entity Recogni-
tion, in: Proceedings of CoNLL-2002/Roth, Dan [edit.], 2002,
pp. 155–158.
[15] S. Pradhan, A. Moschiti, N. Xue, H.T. Ng, A. Bjørkelund,
O. Uryupina, Y. Zhang and Z. Zhong, Towards robust linguis-
tic analysis using OntoNotes, in: Proceedings of the Seven-
N. Jain et al. / Overcoming Shortage of Training Data for Cultural Heritage Analysis 13

(10) (2014), 78–85.


[71] SpaCy, version 2.1.3.