Semantic Web 0 (0) 1
IOS Press

Generation of Training Data for Named Entity Recognition of Artworks

Nitisha Jain, * Alejandro Sierra-Múnera, Jan Ehmueller, Ralf Krestel
Hasso Plattner Institute, University of Potsdam, Germany
E-mails: Nitisha.Jain@hpi.de, Alejandro.Sierra@hpi.de, Jan.Ehmueller@student.hpi.uni-potsdam.de, Ralf.Krestel@hpi.de

Editors: Mehwish Alam, FIZ Karlsruhe, Germany; Victor de Boer, Vrije Universiteit Amsterdam, Netherlands; Enrico Daga, The Open University, United Kingdom; Marieke van Erp, KNAW Humanities Cluster, Netherlands; Eero Hyvönen, University of Helsinki, Aalto University, Finland; Albert Meroño Peñuela, Vrije Universiteit Amsterdam, Netherlands; Harald Sack, FIZ Karlsruhe, Germany
Solicited reviews: Marijn Koolen, Huygens Institute for the History of the Netherlands, Netherlands; Marieke van Erp, KNAW Humanities Cluster, Netherlands; Two anonymous reviewers

Abstract. As machine learning techniques are being increasingly employed for text processing tasks, the need for training data has become a major bottleneck for their application. Manual generation of large scale training datasets tailored to each task is a time consuming and expensive process, which necessitates their automated generation. In this work, we turn our attention towards creation of training datasets for named entity recognition (NER) in the context of the 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 the baseline for NER performance for artwork titles when models are trained on the dataset generated using our framework.

Keywords: Training Data Generation, Named Entity Recognition, Cultural Heritage Data, Weakly-supervised Learning

1. Introduction1

Deep learning models have become popular for natural language processing (NLP) tasks in recent years [1]. 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 [2–4]. As supervised learning techniques have become ubiquitous, the availability of training data has emerged as one of the major challenges for their success [5]. 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 [6–9]. 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 catego-

*Corresponding author.

rize the found named entities into different types. NER serves as an important step for various semantic tasks, such as knowledge base creation [10], machine translation [11], relation extraction [12] and question answering [13]. Most NER efforts are restricted to only 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 [14, 15].

FiNER helps to precisely determine 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 [14, 16]. 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 [17, 18]. Fine-grained NER is also essential for domain-specific NER, where 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 *Banks*, *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 CoNNL datasets [7, 19] or OntoNotes [20]. 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 [21, 22].

This is especially true for cultural heritage data where the cultural artefacts serve as one of the most important named entity categories. Recently, there has been a surge in the availability of digitized fine-arts collections with the principles of linked open data[1] gaining momentum in the cultural heritage domain [23]. The semantic web plays a central role as the enabler of these technologies for the sharing and linking of data between various GLAM(Galleries, Libraries, Archives and Museums) institutions across the world [24, 25]. Initiatives such as OpenGLAM[2] and flagship digital library projects such as Europeana[3] aim to enrich open knowledge graphs with cultural heritage data by improving the coverage of the topics related to the cultural domain. In this direction, efforts have been made to digitize historical archives in various domains. Particularly in the art domain, a large collection of raw texts are yet to be explored and analysed. These collections consist of fine-arts related texts such as auction catalogues, art books and exhibition catalogues [26, 27]. In such resources, cultural objects, mainly artworks such as paintings and sculptures, 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[4].

While several previous works on FiNER have defined entity types ranging from hundreds [14, 15, 29] to thousands [30] of different types, they are not specifically catered to the art domain. Ling et al. [14] have defined 112 named entity types from generic areas. Similarly to Gillick et al. [15], 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

[1]Linked Open Data: http://www.w3.org/DesignIssues/LinkedData

[2]OpenGLAM: http://openglam.org

[3]Europeana: http://europeana.eu

[4]Note that in this work, we use the term 'artworks' to primarily refer to fine-arts such as paintings and sculptures that are dominant in the digitized collections that constitute our dataset. This term is inspired from previous related work such as [28].

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[5]. To this end, we leverage existing art resources that are integrated in popular knowledge bases, such as Wikidata [31] and the Getty vocabularies [32] to first create entity dictionaries for matching and tagging artwork titles. We also incorporate entity and dataset labelling functions with the help of the Snorkel system [33] to learn useful patterns for annotating training data. 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 for two popular NER models 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. Note that while we focus on paintings and sculptures in this work (that reflect the dominant artworks in our dataset), the proposed techniques can be adapted and expanded for other general forms of artworks such as music, novels, films and video games as well, however this is beyond the scope of the present work.

This work was first introduced in Jain et al. [34]. We have since significantly extended the techniques for the generation of the training data, that has enabled us to report better NER performance in this version. Specifically, we have made the following additional contributions — The introduction section includes a discussion with respect to existing efforts about the limitations of OCR quality when it comes to digitization of old cultural resources and the challenges it poses for the performance of natural language processing tools for such corpora. The related work section has been expanded to include the recent works and a subsection to discuss and compare the previous works that have leveraged the Wikipedia texts for NER similar to our work has been added. Section 3 presents an exploration of the unique issues for the identification of artwork titles from a linguistic perspective and the errors that arise as a consequence of the linguistic phenomena. We have significantly extended our approach for the generation of training data by expanding the entity dictionaries and leveraging the Snorkel system for incorporating labelling functions for annotations. Further, recognizing the limitations of the quality of the training data due to a noisy underlying corpus, we attempt to get clean and well-structured texts from existing available resources (such as Wikipedia) to generate silver-standard training data. The resulting improvement in performance justifies the efficacy of the approach. In the experiments, a second baseline NER model has been added to strengthen the evaluation. Furthermore, a detailed error analysis and discussion of the results of the semi-automated approach has been added. The last section introduces the first version of our NER demo that illustrates the results of our approach and enables user interaction.

The rest of the paper is organized as follows — In the next section, we compare and contrast the research efforts related to our work. Section 3 elaborates on the specific challenges of NER for artworks to motivate the problem. In Section 4, we describe our approach to tackle these challenges and generate large corpus of labelled training data for identification of titles. In Section 5, we explain the experimental setup and present the results of our evaluation. Section 6 provides an analysis and further discussion of the results. Finally, Section 7 provides a glimpse of our demo that illustrates the NER performance for artwork titles through an interactive and user-friendly interface.

2. Related Work1

We discuss the related work under different categories, starting 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,

[5]NER is a language specific task and we focus in this work on the English language texts that constitute a majority in our dataset.

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) [35–38]. In the past decade, such systems have been succeeded by neural network based architectures that do not rely on hand-crafted features to identify named entities correctly. Many architectures leveraging Recurrent Neural Networks (RNN) for word level representation [39–41], and Convolutional Neural Networks (CNN) for character level representation [42–44] 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 CoNNL 2002 and 2003 datasets [2, 45, 46]. Additionally, current state-of-the-art NER approaches make use of pre-trained embedding models, both on word and character level, as well as language models and contextualized word embeddings [47–49].

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 CoNNL 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. Also, these datasets include a predefined set of named entity categories, which might not correspond in different entity domains. In most cases, these systems fail to adapt well to new domains and different named entity categories [21, 22].

2.1. Domain specific NER.2

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 [50–52]. But since these techniques rely on specific resources such as carefully curated lists for drug names [53] or biology and microbiology NER datasets [54, 55], 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. [56] 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. [57] 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, Rodriquez et al. [58] 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. [59] offer a diachronic evaluation of various NER tools for digitized archives of Swiss newspapers. Freire et al. [60] use a CRF-based model to identify persons, locations and organizations on cultural heritage structured data. However, none of the existing works have focused on the task of identifying titles of paintings and sculptures 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 for cultural heritage, whereas in this work we aim to improve the performance of NER systems by generating domain-specific high-quality training data. In the context of online book discussion forums, there are few efforts to identify and link the mentions of books and authors [61–63]. While this work is related to ours since books can also be considered as part of cultural heritage, previous work has relied primarily on manually generated annotations and supervised techniques. Such techniques are not scalable to other entity types due to the lack of reliable annotations for training purpose. Recently, there has been increasing effort to publish cultural heritage collections as linked data [27, 64, 65], 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.

2.2. Training Data Generation.2

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 [66], the authors attempt to aid the creation of labeled training data in weakly-supervised fashion by a heuristic based approach. Other works that depend on heuristic patterns along with user input are [67, 68]. In this work, we take the aid of Snorkel [33] for the creation of good quality annotations(Section 4.2). Similar to our approach, Mints et al. [12] leveraged Freebase knowledge base and used distant supervision for training relation extractors. Likewise, Tuerker et al. [69] generate a weakly supervised dataset for text classification based on three embedding models. Two of these models leverage Wikipedia's anchor texts as entity dictionaries with the goal of assigning labels to documents, similar to the manner in which we generate the silver-standard training data for entity recognition.

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 [70, 71]. Ghaddar and Langlais further extended this work by adding more annotations from Wikipedia in [72] and adding fine-grained types for the entities in [73]. 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. Previous works on fine-grained NER have used a generic and cleanly formatted text like Wikipedia to annotate many different entity types. Our focus in this work 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 a 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.

In the next section, we discuss the specific challenges of identifying artwork titles and motivate the necessity of generation of training data for this problem.

3. Challenges for Detecting Artwork Titles1

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* that refers to the most relevant and dominant artworks in our dataset of digitized collections, i.e. paintings and sculptures[6]. Artworks in fine-art collections 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 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 [74]. 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 ex-

[6]The label *artwork* for the new named entity type can be replaced with another such as *fine-art* or *visual-art* without affecting the proposed technique.

Table 1

Types of documents in WPI dataset

Document Type	Count	Ratio
Auction Catalogues	71,192	0.45
Books	42,370	0.27
Exhibition Catalogues	38,176	0.24
Others	7,054	0.04

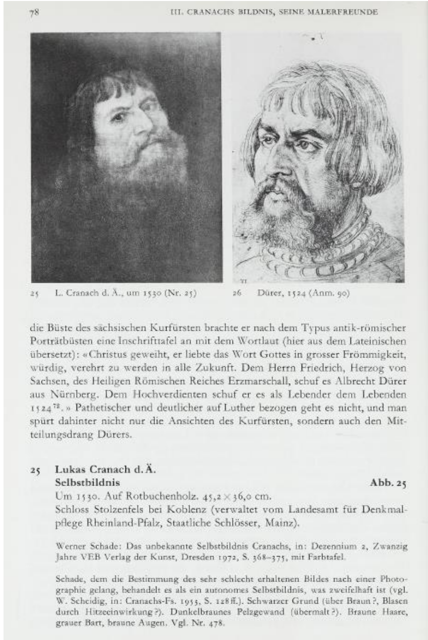
acerbates the challenges for automated text analysis, including the NER task [58].

For this work, the underlying dataset is a large collection of recently digitized art historical documents provided to us by the Wildenstein Plattner Institute (WPI)[7], that was founded to promote scholarly research on cultural heritage collections. This corpus 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. Table 1 shows the proportion of the different kinds of documents in the dataset. For reference, a sample document[8] from a similar collection is shown in Fig. 1. The pages of the catalogues and books in the WPI dataset 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. Fig. 2 shows a typical text excerpt that highlights the noise in the dataset. After OCR of the page, the page numbers are merged with the text, any formatting indicators present in the original page are lost, there are several spelling errors and it is hard to distinguish the artwork title from its description.

In order to systematically highlight the difficulties that arise when trying to recognize artwork mentions in practice, we categorize and discuss the different types of errors that are commonly encountered as follows — 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. Further, there are also errors due to nested named entities and other ambiguities.

[7] https://wpi.art/

[8] from an exhibition catalogue - Lukas Cranach: Gemälde, Zeichnungen, Druckgraphik ; Ausstellung im Kunstmuseum Basel 15. Juni bis 8. September 1974, (https://digi.ub.uni-heidelberg.de/diglit/koepplin1974bd1/0084/image)



78 III. CRANACHS BILDNIS, SEINE MALERFREUNDE

25 L. Cranach d. Ä., um 1530 (Nr. 25) 26 Dürer, 1524 (Anm. 90)

die Büste des sächsischen Kurfürsten brachte er nach dem Typus antik-römischer Porträtbüsten eine Inschrifttafel an mit dem Wortlaut (hier aus dem Lateinischen übersetzt): «Christus geweiht, er liebte das Wort Gottes in grosser Frömmigkeit, würdig, verehrt zu werden in alle Zukunft. Dem Herrn Friedrich, Herzog von Sachsen, des Heiligen Römischen Reiches Erzmarschall, schuf es Albrecht Dürer aus Nürnberg. Dem Hochverdienten schuf er es als Lebender dem Lebenden 1524[72].» Pathetischer und deutlicher auf Luther bezogen geht es nicht, und man spürt dahinter nicht nur die Ansichten des Kurfürsten, sondern auch den Mitteilungsdrang Dürers.

25 **Lukas Cranach d. Ä.**
Selbstbildnis **Abb. 25**
Um 1530. Auf Rotbuchenholz. 45,2 × 36,0 cm.
Schloss Stolzenfels bei Koblenz (verwaltet vom Landesamt für Denkmalpflege Rheinland-Pfalz, Staatliche Schlösser, Mainz).

Werner Schade: Das unbekannte Selbstbildnis Cranachs, in: Dezennium 2, Zwanzig Jahre VEB Verlag der Kunst, Dresden 1972, S. 368–375, mit Farbtafel.

Schade, dem die Bestimmung des sehr schlecht erhaltenen Bildes nach einer Photographie gelang, behandelt es als ein autonomes Selbstbildnis, was zweifelhaft ist (vgl. W. Scheidig, in: Cranachs-Fs. 1955, S. 128 ff.). Schwarzer Grund (über Braun?, Blasen durch Hitzeeinwirkung?). Dunkelbraunes Pelzgewand (übermalt?). Braune Haare, grauer Bart, braune Augen. Vgl. Nr. 478.

Fig. 1. Example of scanned page

```
54
The Property of THE SNITE MUSEUM OF ART, UNIVERSITY OF NOTRE
DAME, INDIANA
78 FLEMISH SCHOOL, 16th Century
Portrait of a gentleman, half length, wearing a dark costume
and a cap
oil on pane1
I2V4 X lOin. (31.1 X 25.5cm.)
S2,000-3,000
Portrat of a young man bust length, wearing a blue vest,
black coat, and a cap, in a landscape
oil 011 panel
16 X 12'/2in. (40.5 x 31.7cm.)
PROVENANCE
Jack Linsky S2,000-3,000
80
```

Fig. 2. Example of digitized text

3.1. Incorrectly Missed Artwork Title2

Many artwork titles contain generic words that can be found in a 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.

3.2. Incorrect Artwork Title Boundary Detection2

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.

3.3. Incorrect Type Tagging of Artwork Title2

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*.

3.4. Nested Named Entities2

Yet another type of ambiguity involving both incorrect boundaries and wrong tagging can occur in the context of nested named entities, where 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 by the default SpaCy NER tool. Due to the often descriptive nature of artwork titles, it is quite common to encounter *person* or *location* named entities embedded within the artwork titles which lead to confusion and errors in the detection of the correct *artwork* entity. Therefore, careful and correct boundary detection for the entities is imperative for good performance[9].

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 as already illustrated in Fig. 2 that cannot be eliminated without manual efforts. 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. We discuss in detail our approach for generating annotations for NER from a large corpus of art related documents in the next section.

4. Generating Training Data for Artwork Titles1

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. 3). 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 and refine annotations for *artwork* named entities. A set of training datasets is obtained at the end of each stage, with the final annotated dataset

[9]Details on how our approach handles this complexity are presented in Section 4.1

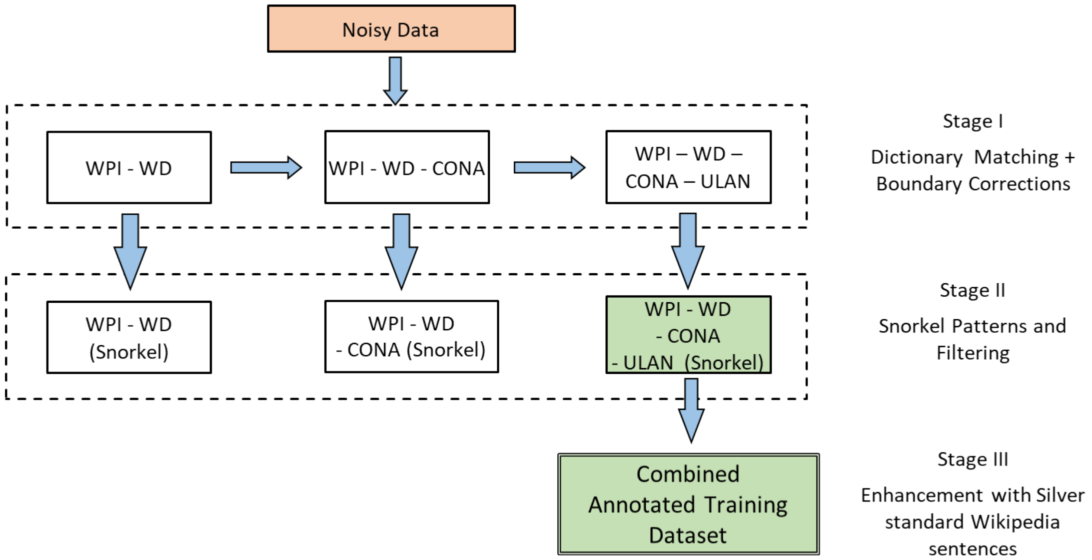


Fig. 3. Overview of the framework showing the progressive improvements of the training datasets (as described in Section 4 and summarized in Table 2). Each stage illustrates the enhancements by referring to the datasets obtained with the corresponding steps.

being the best performing version. While the artwork titles are multi-lingual, we focus on English texts in this work and plan to extend to further languages in future efforts. We describe the three stages of the framework and the output datasets at each stage.

4.1. Stage I - Dictionary-based matching for labelling artwork titles2

In the first stage, we aimed to match and correctly tag the artworks present in our corpus as named entities with the help of entity dictionaries to obtain highly precise annotations. Apart from extracting the existing artwork titles from the structured part of the WPI dataset (1,075 in total), we leveraged other cultural resources that have been integrated into the public knowledge bases such as Wikidata, as well as linked open data resources such as the Getty vocabularies for creating these dictionaries. As a first step, we collected available resources from Wikidata to generate a large entity dictionary or *gazetteer* of artwork titles in an automatic way. To generate the entity dictionary for titles, Wikidata was queried with the Wikidata Query Service[10] 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 ...* ’). It was noticed that quite a few of the titles having only one word were highly generic, for instance, ‘*Italian*’, ‘*Winter*’, ‘*Landscape*’, ‘*Portrait*’ etc. Matching with such titles was contributing to errors in the annotation process, since common words in the description of the artworks were being wrongly tagged as the *artwork* named entity. In order to maintain high precision of annotations in the first stage, the titles having only one word were removed from the list even at the slight expense of missed tags for some valid artwork titles. Since several artwork titles are identical to location names such as ‘*Germania*’, ‘*Olympia*’ which can lead to errors while tagging the named entity to the correct type, such titles were also ignored. Overall, around 5% of the titles were removed in this manner[11]. A combined list of approximately 15,000 titles in different languages was obtained, the majority of the ti-

[10]https://query.wikidata.org/

[11]one-word titles are encountered during training in Stage III.

tles being in English. 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. 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 annotations obtained from the combined WPI and Wikidata entity dictionary resulted in the first version of the training dataset, referred to as *WPI-WD*.

Furthermore, we explored the Getty vocabularies, such as CONA and ULAN, that contain structured and hand-curated terminology for the cultural heritage domain and are designed to facilitate shared research for digital art resources. The Cultural Objects Named Authority (CONA) vocabulary[12] comprises titles of works of art and architecture. Since these are contributed and compiled by an expert user community, these titles are highly precise and can lead to good quality annotations. A total of 3,013 CONA titles were added to the entity dictionary. The Union List of Artist Names (ULAN)[13] contains names of artists, architects, studios and other bodies. We mainly extracted artist names from this list (899,758 in total) and tagged them in our corpus via matching, with the motivation of providing additional context for the identification of artwork titles through pattern learning. Different versions of the dataset were generated after the iterative enhancements in annotations by the use of CONA titles and ULAN names, referred to as *WPI-WD-CONA* and *WPI-WD-CONA-ULAN* respectively.

In all cases, 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 5.2). 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.

Improving Named Entity Boundaries.4 As discussed in Section 3.2, 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 for each of the datasets obtained by dictionary matching. For a given text, a separate list of matches with the artwork titles in the entity dictionary over the entire text were 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, consider the nested named entity 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, thus resolving the particularly challenging issue of missing or wrong tagging for nested named entities. Through this technique, many missed mentions of artwork titles were added to the training datasets generated in this stage, thus improving the recall of the annotations and the overall quality of the datasets.

4.2. Stage II - Filtering with Snorkel Labelling Functions2

Identification of artwork titles as named entities from unstructured and semi-structured text can be aided with the help of patterns found in the text. To leverage these patterns, we use Snorkel, an open source system that enables the training of models without hand labeling the training data [33] with the help of

[12] Getty CONA (2017), http://www.getty.edu/research/tools/vocabularies/cona, accessed October 2020.

[13] Getty ULAN (2017), http://www.getty.edu/research/tools/vocabularies/ulan, accessed October 2020.

Fig. 4. Getting annotated sentences from Wikipedia

a set of labelling functions and patterns. It combines user-written labelling functions and learns their quality without access to ground truth data. Using heuristics, Snorkel is able to estimate which labelling functions provide high or low quality labels and combines these decisions to a final label for every sentence. This functionality is used for deciding whether an annotated sentence is of high-quality, such that it is retained in the training data while the low-quality sentences can be filtered out. Since the training dataset contained a number of noisy sentences that are detrimental to model training, Snorkel helped in reducing the noise by identifying and filtering out these sentences, while at the same time increasing the quality of the training data.

Based on the characteristics of the training data, a set of seven labelling functions were defined to capture observed patterns. For example, one such labelling function expresses that a sentence is of high-quality if it contains the phrase "*attributed to*" that is preceded by a *artwork* annotation and also succeeded by a *person* annotation. This pattern matches many sentences containing painting descriptions in auction catalogues, which make up a large part of our dataset. Another labelling function expresses that a sentence is a low-quality sentence, if it contains less than 5 tokens. With this pattern many noisy sentences are removed that were created either by OCR errors as described in Section 3 or by sentence splitting errors that were caused due to erroneous punctuation. By only retaining the sentences that are labeled as high-quality by Snorkel, the amount of training data is drastically reduced, as can be seen in Table 2. The resulting datasets include annotations of higher quality that can be used to more efficiently train an NER model while reducing the noise. As an example, in the case of the WPI-WD dataset (that contains annotations obtained from matching titles in the combined entity list from WPI titles and Wikidata titles), using Snorkel reduces the number of sentences to 3.2% of the original size, while only reducing the number of artwork annotations to 25.5% of the previous number.

Table 2

Statistics of datasets

Training Dataset	Sentences	Annotations	Unique entities
Ontonotes5	185,254	1,650	-
WPI-WD	13,383,185	1,933,119	36,720
WPI-WD-CONA	13,383,185	1,951,070	37,271
WPI-WD-CONA-ULAN	13,383,185	1,875,711	36,715
WPI-WD (Snorkel)	437,026	492,192	21,838
WPI-WD-CONA (Snorkel)	436,953	496,591	22,027
WPI-WD-CONA-ULAN (Snorkel)	433,154	482,562	21,684
Wikipedia	1,628	1,835	587
Combined Annotated Dataset	434,782	484,397	22,271

At the end of this stage, we obtained high-quality, shrunk down versions of all three training datasets that led to improved performance of the NER models trained on them.

4.3. Stage III - Enhancements with Silver Standard Training Data2

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 further learn the textual indicators present in the dataset for identification of artworks, in this stage we augmented our best performing training dataset with clean and well-structured silver standard[14] 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 2,808 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 structure 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 [75–77]. We followed a somewhat similar approach — for each Wikipedia page referring to 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 1,628 sentences were added as silver standard annotation data to the training set. The process is illustrated in Fig. 4. This data provided correct and precise textual patterns that were highly indicative of the artwork titles and led to a considerable boost in training data quality. This dataset was augmented to the best performing dataset obtained from the previous stages (*WPI-WD-CONA-ULAN (Snorkel)*) to generate a combined annotated dataset as the final result of the framework. It is to be noted that at this stage, artwork titles having a single word were also included in the annotations such that the trained model could learn from them. Overall, the final model is expected to show a more favorable performance towards multi-word titles. However, the number of false positives for one-word titles would be lower due to high quality annotations from the silver-standard annotations.

[14]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.

5. Evaluation and Results1

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

5.1. Experimental Setup2

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 removing all non-English texts and performing initial preprocessing, including the removal of erroneous characters, the dataset included both partial sentences such as artwork size related entries as well as well-formed sentences describing the artworks. This 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 two well-known machine learning based NER models, SpaCy and Flair, for the new entity type *artwork* on different variants of training data as shown in Table 2 and measured their performance.

5.2. Baselines2

None of the existing NER systems can identify titles of artworks as named entities out-of-the-box. While previous works such as [15] and [14] consider a broad 'art' entity type, they do not include paintings and sculptures which are the primary focus of this work. Thus, these could not serve as baselines for comparison. The closest NER category to artwork titles was found in the Ontonotes5 dataset[15] 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. In this work, we seek to perform NER for a particular subset of this category, i.e. paintings and sculptures. Therefore, we aim to train the NER models to perform the complex task of learning the features for paintings and sculptures, while at the same time separating them from other cultural heritage objects

[15]https://catalog.ldc.upenn.edu/LDC2013T19

such as book, music etc.[16] For the lack of alternatives, we have leveraged the *work_of_art* NER category in our work for setting up a naive baseline in which the training was performed on more general annotations. With this baseline, we will compare the improvements in NER performance obtained by retraining the tools on our semi-automatically generated corpus with the specialized *artwork* entity type.

To quantify the performance gains from annotations obtained at each stage, SpaCy and Flair NER models were re-trained on each of the generated datasets for a limited number of epochs (as per computational constraints), with the training data batched and shuffled before every iteration. In each case, the performance of the re-trained NER models was compared with the *baseline* NER model (the pre-trained model without any specific annotations for artwork titles). As the underlying Ontonotes dataset does not have *artwork* annotations, the named entity type *artwork* was not applicable for the baseline models of SpaCy and Flair. Therefore, a match with the entity type *work_of_art* was considered as a true positive during the evaluations. 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.

SpaCy.4 The SpaCy[17] library is popular for many natural language processing tasks including named entity recognition. SpaCy text processing tools were 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 the Ontonotes5 dataset 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*. With the pre-trained SpaCy NER model as baseline, the model was trained on the datasets for 10 epochs each and the performance evaluated.

Flair.4 Similar to SpaCy, Flair [78] is another widely used deep-learning based NLP library that provides an NER framework in the form of a sequence tagger, pre-trained with the Ontonotes5 dataset. The best configuration reported by the authors for the Ontonotes dataset, was re-trained with a limited number of epochs in order to define a baseline to compare against the datasets proposed in this paper. The architecture of the sequence tagger for the baseline was configured to use stacked GloVe and Flair forward and backward embeddings [49, 79]. For training the model the following values were assigned to the tagger hyper-parameters: learning rate was set to 0.1, and the number of epochs was limited to 10. These values and the network architecture were kept throughout all the experiments in order to achieve a fair comparison among the training sets.

It is to be noted that the techniques for improving the quality of NER training data that are proposed in this work are independent of the NER model used for the evaluation. Thus, SpaCy and Flair can be substituted with other re-trainable NER systems.

5.3. Manual Annotations for Test Dataset2

To generate 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 such that there was no entity overlap between the two. The titles of paintings and sculptures mentioned in this data were manually identified and tagged as named entities of *artwork* type. The annotations were performed by two non-expert annotators (from among the authors) independently of each other in $3-4$ person hours with the help of the Enno[18] 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 3) 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;*

[16] Further analysis with examples is presented in Section 6.2 while discussing the results.

[17] SpaCy: https://spacy.io/

[18] https://github.com/HPI-Information-Systems/enno

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*'. It was difficult to correctly tag these artwork mentions without having expert knowledge of the art domain, especially with regard to the particular period of art. Due to these reasons, the inter-annotator agreement was quite low. The Fleiss' kappa [80] and Krippendorff's alpha [81] scores were calculated as -1.86 and 0.61 respectively. (A negative Fleiss' kappa score indicates poor agreement, while Krippendorff's alpha values for data should be above 0.667 to be considered useful). The poor inter-annotator agreement reflected by these scores reaffirmed that the task of annotating the artwork titles is difficult, even for humans. Only experts in the particular artwork collections could have perhaps identified the artworks correctly, however such expertise is rarely available or even practical. Therefore, in order to obtain the gold standard test dataset for the evaluation of NER models, the disagreements were manually sorted out with the help of web search to the best of our understanding and a total of 144 entities were positively tagged as *artwork*.

5.4. Evaluation Metrics2

The performance of NER systems is generally measured in terms of precision, recall and F1 scores. The correct matching of a named entity involves the matching of the boundaries of the entity (in terms of character offsets in text) as well as the tagging of the named entity to the correct category. The strict F1 scores for NER evaluation were used in the CoNNL 2003 shared task[19], where the entities' boundaries were matched exactly. The MUC NER task[20] allowed for relaxed 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* entities 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 3.2, we evaluated the NER models with both strict metrics based on exact boundary match, as well as the relaxed metrics based on partial boundary matches. The relaxed F1 metric allowed for comparison of the entities despite errors due to wrong chunking of the named entities in the text. Precision, recall, as well as F1 scores obtained for the NER models trained with different training dataset variants are shown in Table 3.

6. Analysis and Discussion1

The results demonstrated definitive improvement in performance for the NER models that were trained with annotated data as compared to the baseline performance. Since the relaxed metrics allowed for flexible matching of the boundaries of the identified titles, they were consistently better than the strict matching scores for all cases. The training data obtained from Stage I, i.e. the dictionary based matching, enabled an improvement in NER performance due to the benefit of domain-specific and entity-specific annotations generated from the Wikidata entity dictionaries and Getty vocabularies, along with the boost from additional annotations by the correction of entity boundaries. Further, the refinement of the training datasets obtained with the help of Snorkel labelling functions in Stage II led to better training of the NER models reflecting in their higher performance especially in terms of recall. To gauge the benefits from the silver standard annotations from Wikipedia sentences, a model was trained only on these sentences(Stage III). It can be seen that the performance of this model was quite high despite the small size of the dataset, indicating the positive impact of the quality of the annotations. The NER models re-trained on the combined annotated training dataset obtained through our framework, consisting of all the annotations obtained from the three stages, showed the best overall performance with significant improvement across all metrics, particularly in terms of recall. This indicates that the models were able to maintain the precision of the baseline while being able to find much more entities in the test dataset. The encouraging results demonstrate the importance 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

[19]https://www.clips.uantwerpen.be/conll2003/ner/

[20]https://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html

Table 3
Performance of NER Model trained on different datasets

Training Dataset	Stage	*SpaCy*						*Flair*					
		Strict			*Relaxed*			*Strict*			*Relaxed*		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Default Unannotated (baseline)	–	.14	.06	.08	.22	.08	.12	**.22**	.04	.07	**.29**	.05	.09
WPI-WD	I	.24	.23	.23	.41	.42	.41	.03	.05	.04	.06	.09	.07
WPI-WD-CONA	I	.27	.26	.26	.43	.45	.44	.04	.08	.06	.08	.14	.10
WPI-WD-CONA-ULAN	I	.28	.26	.27	.48	.45	.46	.05	.08	.07	.09	.14	.11
WPI-WD (Snorkel)	II	.31	.28	.30	.50	.49	.50	.07	.12	.08	.12	.21	.15
WPI-WD-CONA (Snorkel)	II	.31	.31	.31	.53	.51	.52	.07	.11	.08	.13	.22	.17
WPI-WD-CONA-ULAN (Snorkel)	II	.32	.33	.33	.55	.51	.53	.09	.16	.11	.14	.24	.18
Wikipedia	III	.17	.13	.15	.38	.30	.34	.12	.34	.17	.21	**.61**	.31
Combined Annotated Dataset	All	**.46**	**.41**	**.43**	**.68**	**.62**	**.65**	.21	**.45**	**.29**	.28	.59	**.38**

Table 4
Performance of NER models trained on different dataset sizes

Dataset Size	*SpaCy*						*Flair*					
	Strict			*Relaxed*			*Strict*			*Relaxed*		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
5%	.17	.12	.14	.24	.37	.30	.08	.12	.09	.23	.34	.27
10%	.24	.16	.20	.27	.40	.33	.12	.23	.16	.21	.43	.28
15%	.27	.28	.27	.35	.38	.36	.18	.37	.24	.24	.48	.32
20%	.31	.30	.30	.36	.40	.38	.15	.29	.20	.24	.45	.32
25%	.32	.34	.33	.42	.45	.43	.16	.27	.20	.25	.42	.31
50%	.36	.39	.37	.48	.55	.51	.17	.40	.24	.24	.57	.34
75%	.39	.38	.39	.55	.54	.55	.15	.29	.20	.27	.51	.35
100%	.46	.41	.43	.68	.62	.65	.21	.45	.29	.28	.59	.38

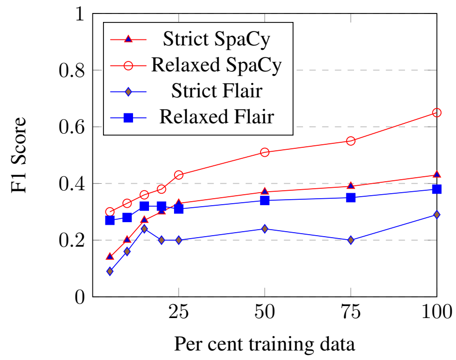


Fig. 5. NER performance with different training data sizes

entity of type *artwork*, proves the effectiveness of our approach.

We have released[21] the annotated datasets as well as the NER models trained on these datasets for the benefit of the research community and to foster further efforts in this direction. Since the annotations in this dataset have been derived from public datasets such as Getty and Wikipedia, it is expected that the annotations would lean more towards the popular artwork names that were found in these resources i.e. there might be a certain skew towards the popular mentions and bias against the less known artwork names. However, modern NER models including SpaCy and Flair that are trained on this dataset would learn to identify the textual patterns related to the labeled annotations on artworks, as discussed earlier. As such, once these models are trained, having learnt the useful features during the training, they should be able to identify the mention of any artwork, whether famous or less-known, merely based on the context. Thus, in our view the NER models trained on these datasets are not expected to have any significant bias towards the popular artworks and would be useful for both domain experts and non-experts.

In the remainder of this section we study the impact of the size of the training data on the models' performance, as well as present a detailed discussion on error analysis.

6.1. *Impact of Training Data Size2*

To inspect the effect of the size of the generated training data on NER performance, we varied the

[21]https://github.com/HPI-Information-Systems/art-ner-dataset

Table 5
Analysis of extracted artwork titles

#	Text with true title underlined	Extracted title	Category
1	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.	On the Terrace	True Positive
2	. . . as in End of a Gambling Quarrel (Fig. 45), where the furniture is overturned, one chair projecting to the very picture surface, and the cards are strewn . . .	End of a Gambling Quarrel	True Positive
3	He owned a painting entitled The Little Nephew of Rameau (1858), a rare instance of Meissonier making a literary allusion.	The Little Nephew of Rameau	True Positive
4	Figure 34. The Inn Door in the Saint-Germain Forest. Panel, 17 x 23 cm. Paris, Musee d'Orsay.	The Inn Door	Partial Match
5	Among the other works in Davis's private collection was The Grand Canal with Ca' Pesaro by Francesco Guardi, sold at Christie's, London.	The Grand Canal	Partial Match
6	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 . . .	The Three Musketeers	False Positive
7	Property from the Collection of William And Eleanor Wood Prince, CHICAGO, ILLINOIS	William And Eleanor Wood Prince	False Positive
8	. . . 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).	—	False Negative
9	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.	—	False Negative

dataset size and performed the model training on progressively increasing sizes of training data. We randomly sampled smaller sets from the overall training dataset in the range 5 per cent to 100 per cent and plotted the performance scores of the trained models (averaged over 10 iterations) as shown in Fig. 5. The detailed scores are shown in Table 4. It can be seen that all the scores show a general upward trend as the training data size increases. Initially, the numbers get rapidly better with increasing training data sizes and then stabilize over time with smaller gains. It is interesting to note that the SpaCy model continues to show improvement under the relaxed setting, suggesting that the model gets better at identifying the titles, but not with the exactly correct boundaries. The overall best scores were achieved with the entire training dataset that was obtained as output from the framework. This suggests that if the training dataset is further enlarged, the performance of the models trained with it will likely improve.

6.2. Error Analysis2

A closer inspection of the performance of NER models revealed interesting insights. Some example annotations performed by the best-trained Spacy NER model are shown in Table 5. As discussed in Section 3, 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 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 the 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 5. Given the particular use case of noisy art collections and the ambiguities inherent in artwork titles, this is indeed a hard problem to tackle. Similar boundary errors were also made by the human annotators. The relaxed metrics consider partial matches as positive matches and favour the trained NER model in such cases.

There were also a few interesting 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 Text 6. Due to the fact that books and other cultural objects such as plays, films, music often occur in similar contexts, the NER model finds it particularly hard to separate mentions of paintings and sculptures from the other types of artwork mentions. This is indeed a challenging problem that could likely be solved only with manual efforts by domain experts to obtain gold standard annotations for training. In some cases, the names of persons 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 detect the titles of artworks due to lack of hints or familiar patterns to rely upon. In some cases such mentions were indeed hard to identify even during the manual annotations and further exploration had been needed for correct tagging. In spite of the difficulties for this specific entity type, it is encouraging to note the improvement of performance of the NER model, making the case for the usefulness of the generated training data by our framework.

7. NER Demo1

We present an on-going effort to build an end user interface[22] that demonstrates the performance of our proposed approach. The best performing NER model obtained by re-training on the combined annotated training dataset is used for annotating named entities, including the *artwork* type on sample texts. The demo includes a few example texts and can also take user provided texts as input. This system comprises two components: a front-end graphical interface for facilitating user interaction and an annotation service at the back-end that provides the output from the NER model to be displayed to the user.

[22] https://hpi.de/naumann/sites/wpi_demo/demo

Fig. 6 shows the user interface of this system. On the left is the text area in which an example text is displayed. This is also where a user can edit or paste any texts that need to be annotated. The named entity tags are then fetched from the trained models at the back-end, the user can choose to display the results from the Flair model or the default SpaCy model. After the results are fetched, all the identified named entities in the text are highlighted by their respective type of named entity labels. The labels are explained on the right and highlighted with different colors for clarity and easy identification of the entity types. The demo can be explored with a few sample texts from the drop down menu, which will be annotated upon selection. Additionally, a user can click on any label to hide and unhide named entities belonging to this label. We plan to further enhance this NER demo by enabling users to upload text files and integrating named entity linking and relation extraction features in the near future.

8. Conclusion and Future Work1

In this work we proposed a framework to generate a large number of annotations for identifying artwork mentions from art collections. We motivated the need for NER training on high-quality annotations and proposed techniques for generating the relevant training data for this task in a semi-automated manner. Experimental evaluations showed that the NER performance 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 also be adapted for finding fine-grained entity types in other domains, where there is shortage of annotated training data but raw text and dictionary resources are available.

As future work, we would like to apply our techniques for named entity recognition to other important entities such as auctions, exhibitions and art styles to facilitate entity-centric text exploration for cultural heritage resources. 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. The task of named entity linking for artworks is likewise an interesting challenge for future efforts, where the identified artworks would need to

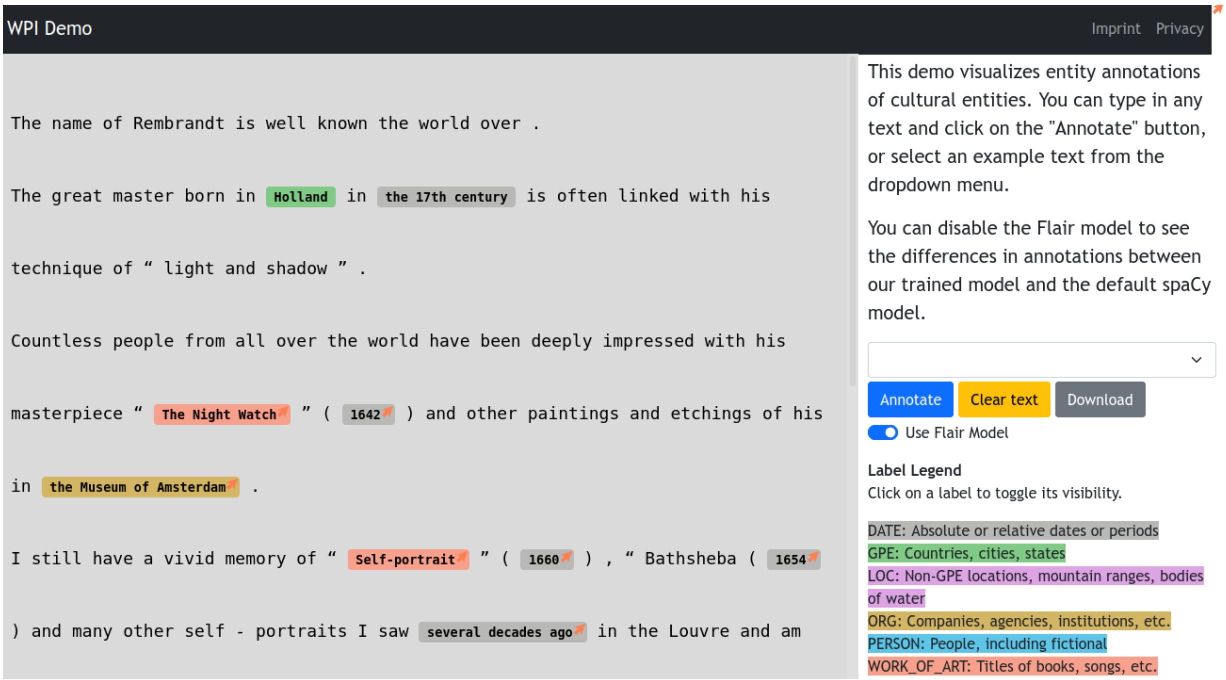


Fig. 6. First version of NER demo system

be mapped to the corresponding instance on existing knowledge graphs. It would be also appealing 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.

Acknowledgements

We thank the Wildenstein Plattner Institute for providing the digitized corpus used in this work. We also thank Philipp Schmidt for the ongoing efforts for the improvement and deployment of the NER demo. This research was partially funded by the HPI Research School on Data Science and Engineering.

References

[1] R. Socher, Y. Bengio and C.D. Manning, Deep learning for NLP (without magic), in: *Tutorial Abstracts of ACL 2012*, 2012, pp. 5–5.

[2] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, 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*, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. doi:10.18653/v1/N16-1030. https://aclanthology.org/N16-1030.

[3] J.P.C. Chiu and E. Nichols, Named Entity Recognition with Bidirectional LSTM-CNNs, *Transactions of the Association for Computational Linguistics* **4**(1) (2016), 357–370.

[4] V. Yadav and S. Bethard, A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, in: *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2145–2158. https://www.aclweb.org/anthology/C18-1182.

[5] C. Sun, A. Shrivastava, S. Singh and A. Gupta, Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[6] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. doi:10.18653/v1/D16-1264. https://aclanthology.org/D16-1264.

[7] E.F. Tjong Kim Sang and F. De Meulder, 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*, 2003, pp. 142–147. https://aclanthology.org/W03-0419.

[8] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A. Ng and C. Potts, Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 1631–1642. https://aclanthology.org/D13-1170.

[9] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng and C. Potts, Learning word vectors for sentiment analysis, in: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, Association for Computational Linguistics, 2011, pp. 142–150.

[10] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, 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 conference on Knowledge discovery and data mining*, ACM, 2014, pp. 601–610.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, 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 computational linguistics companion volume proceedings of the demo and poster sessions*, 2007, pp. 177–180.

[12] M. Mintz, S. Bills, R. Snow and D. Jurafsky, Distant supervision for relation extraction without labeled data, in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, Association for Computational Linguistics, 2009, pp. 1003–1011.

[13] T. Lin, O. Etzioni et al., No noun phrase left behind: Detecting and typing unlinkable entities, in: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, Association for Computational Linguistics, 2012, pp. 893–903.

[14] X. Ling and D.S. Weld, Fine-grained entity recognition, in: *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

[15] D. Gillick, N. Lazic, K. Ganchev, J. Kirchner and D. Huynh, Context-dependent fine-grained entity type tagging, *arXiv preprint arXiv:1412.1820* (2014).

[16] M. Koch, J. Gilmer, S. Soderland and D.S. Weld, Type-aware distantly supervised relation extraction with linked arguments, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1891–1901.

[17] L. Dong, F. Wei, H. Sun, M. Zhou and K. Xu, A hybrid neural model for type classification of entity mentions, in: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015.

[18] X. Li and D. Roth, Learning question classifiers, in: *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, 2002, pp. 1–7.

[19] E.F. Tjong Kim Sang, Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, in: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. https://aclanthology.org/W02-2024.

[20] S. Pradhan, A. Moschitti, N. Xue, H.T. Ng, A. Björkelund, O. Uryupina, Y. Zhang and Z. Zhong, Towards robust linguistic analysis using OntoNotes, in: *Proceedings of the 17th Conference on Computational Natural Language Learning*, 2013, pp. 143–152.

[21] T. Poibeau and L. Kosseim, Proper name extraction from non-journalistic texts, *Language and computers* **37** (2001), 144–157.

[22] R. Prokofyev, G. Demartini and P. Cudré-Mauroux, Effective named entity recognition for idiosyncratic web collections, in: *Proceedings of the 23rd international conference on World Wide Web (WWW)*, ACM, 2014, pp. 397–408.

[23] S. Van Hooland and R. Verborgh, *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*, Facet publishing, 2014.

[24] J. Oomen, M. van Erp and L. Baltussen, Sharing cultural heritage the linked open data way: Why you should sign up, in: *Museums and the Web 2012*, 2012.

[25] A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach and F.V. Harmelen, Semantic technologies for historical research: A survey, *Semantic Web* **6** (2015), 539–564.

[26] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela and S. Kettula, MuseumFinland : Finnish Museums on the Semantic Web, *Web Semantics: Science, Services and Agents on the World Wide Web* (2005), 224–241.

[27] C. Dijkshoorn, L. Jongma, L. Aroyo, J. Van Ossenbruggen, G. Schreiber, W. ter Weele and J. Wielemaker, The Rijksmuseum collection as linked data, *Semantic Web* **9**(2) (2018), 221–230.

[28] B. Saleh and A. Elgammal, Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature, *International Journal for Digital Art History* (2016). doi:10.11588/dah.2016.2.23376. https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/23376.

[29] M.A. Yosef, S. Bauer, J. Hoffart, M. Spaniol and G. Weikum, Hyena: Hierarchical type classification for entity names, in: *Proceedings of COLING 2012: Posters*, 2012, pp. 1361–1370.

[30] E. Choi, O. Levy, Y. Choi and L. Zettlemoyer, Ultra-Fine Entity Typing, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 87–96. doi:10.18653/v1/P18-1009. https://aclanthology.org/P18-1009.

[31] D. Vrandečić and M. Krötzsch, Wikidata: A Free Collaborative Knowledge Base, *Commun. ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

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

[33] A. Ratner, S.H. Bach, H. Ehrenberg, J. Fries, S. Wu and C. Ré, Snorkel: Rapid training data creation with weak supervision, *Proceedings of the VLDB Endowment* **11**(3) (2017), 269–282.

[34] N. Jain and R. Krestel, Who is Mona L.? Identifying mentions of artworks in historical archives, in: *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, Springer, 2019, pp. 115–122.

[35] G. Zhou and J. Su, Named entity recognition using an HMM-based chunk tagger, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 473–480.

[36] R. Malouf, Markov models for language-independent named entity recognition, in: *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, 2002.

[37] Y. Li, K. Bontcheva and H. Cunningham, SVM based learning system for information extraction, in: *International Workshop on Deterministic and Statistical Methods in Machine Learning*, Springer, 2004, pp. 319–339.

[38] R.K. Ando and T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, *Journal of Machine Learning Research* **6**(Nov) (2005), 1817–1853.

[39] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, Natural language processing (almost) from scratch, *Journal of machine learning research* **12**(Aug) (2011), 2493–2537.

[40] Z. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF models for sequence tagging, *arXiv preprint arXiv:1508.01991* (2015).

[41] Y. Shao, C. Hardmeier and J. Nivre, Multilingual named entity recognition using hybrid neural networks, in: *The Sixth Swedish Language Technology Conference (SLTC)*, 2016.

[42] Y. Kim, Y. Jernite, D. Sontag and A.M. Rush, Character-aware neural language models, in: *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[43] O. Kuru, O.A. Can and D. Yuret, Charner: Character-level named entity recognition, in: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 911–921.

[44] D. Gillick, C. Brunk, O. Vinyals and A. Subramanya, Multilingual Language Processing From Bytes, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 1296–1306. doi:10.18653/v1/N16-1155. https://aclanthology.org/N16-1155.

[45] X. Ma and E. Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1064–1074. doi:10.18653/v1/P16-1101. https://www.aclweb.org/anthology/P16-1101.

[46] V. Yadav, R. Sharp and S. Bethard, Deep affix features improve neural named entity recognizers, in: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, 2018, pp. 167–172.

[47] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423.

[48] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep Contextualized Word Representations, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. doi:10.18653/v1/N18-1202. https://www.aclweb.org/anthology/N18-1202.

[49] A. Akbik, D. Blythe and R. Vollgraf, Contextual String Embeddings for Sequence Labeling, in: *Proceedings of the 27th International Conference on Computational Linguistics COLING 2018*, 2018, pp. 1638–1649.

[50] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi and N. Collier, Introduction to the bio-entity recognition task at JNLPBA, in: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, Citeseer, 2004, pp. 70–75.

[51] Ö. Uzuner, Y. Luo and P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, *Journal of the American Medical Informatics Association* **14**(5) (2007), 550–563.

[52] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D.M. Lowe et al., The CHEMDNER corpus of chemicals and drugs and its annotation principles, *Journal of cheminformatics* **7**(1) (2015), S2.

[53] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu et al., DrugBank 3.0: A comprehensive resource for 'omics' research on drugs, *Nucleic acids research* **39**(suppl_1) (2010), D1035–D1041.

[54] L. Hirschman, A. Yeh, C. Blaschke and A. Valencia, Overview of BioCreAtIvE: critical assessment of information extraction for biology, BioMed Central, 2005.

[55] L. Delėger, R. Bossy, E. Chaix, M. Ba, A. Ferrė, P. Bessieres and C. Nėdellec, Overview of the bacteria biotope task at bionlp shared task 2016, in: *Proceedings of the 4th BioNLP shared task workshop*, 2016, pp. 12–22.

[56] S. Van Hooland, M. De Wilde, R. Verborgh, T. Steiner and R. Van de Walle, Exploring entity recognition and disambiguation for cultural heritage collections, *Digital Scholarship in the Humanities* **30**(2) (2013), 262–279.

[57] R. Segers, M. Van Erp, L. Van Der Meij, L. Aroyo, G. Schreiber, B. Wielinga, J. van Ossenbruggen, J. Oomen and G. Jacobs, Hacking history: Automatic historical event extraction for enriching cultural heritage multimedia collections, in: *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP)*, 2011, pp. 26–29.

[58] K.J. Rodriquez, M. Bryant, T. Blanke and M. Luszczynska, Comparison of named entity recognition tools for raw OCR text, in: *Konvens*, 2012, pp. 410–414.

[59] M. Ehrmann, G. Colavizza, Y. Rochat and F. Kaplan, Diachronic Evaluation of NER Systems on Old Newspapers, in: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 2016, pp. 97–107.

[60] N. Freire, J. Borbinha and P. Calado, An Approach for Named Entity Recognition in Poorly Structured Data, in: *The Semantic Web: Research and Applications*, E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 718–732. ISBN 978-3-642-30284-8.

[61] T. Bogers, I. Hendrickx, M. Koolen and S. Verberne, Overview of the SBS 2016 mining track, in: *Conference and Labs of the Evaluation forum*, CEUR Workshop Proceedings, 2016, pp. 1053–1063.

[62] A. Ollagnier, S. Fournier and P. Bellot, Linking Task: Identifying Authors and Book Titles in Verbose Queries., in: *CLEF (Working Notes)*, Citeseer, 2016, pp. 1064–1071.

[63] H. Ziak and R. Kern, KNOW At The Social Book Search Lab 2016 Suggestion Track., in: *CLEF (Working Notes)*, Citeseer, 2016, pp. 1183–1189.

[64] V. De Boer, J. Wielemaker, J. Van Gent, M. Hildebrand, A. Isaac, J. Van Ossenbruggen and G. Schreiber, Supporting linked data production for cultural heritage institutes: The Amsterdam Museum case study, in: *Proceedings of the Extended Semantic Web Conference*, Springer, 2012, pp. 733–747.

[65] P. Szekely, C.A. Knoblock, F. Yang, X. Zhu, E.E. Fink, R. Allen and G. Goodlander, Connecting the Smithsonian American Art Museum to the linked data cloud, in: *Proceedings of the Extended Semantic Web Conference*, Springer, 2013, pp. 593–607.

[66] P. Varma and C. Ré, Snuba: Automating weak supervision to label training data, *Proceedings of the VLDB Endowment* **12**(3) (2018), 223–236.

[67] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: *Proceedings of the 14th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, 1992, pp. 539–545.

[68] R. Bunescu and R. Mooney, Learning to extract relations from the web using minimal supervision, in: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 576–583.

[69] R. Türker, L. Zhang, M. Alam and H. Sack, Weakly Supervised Short Text Categorization Using World Knowledge, in: *Proceedings of the The International Semantic Web Conference (ISWC 2020)*, Springer International Publishing, Cham, 2020, pp. 584–600. ISBN 978-3-030-62419-4.

[70] J. Nothman, Learning named entity recognition from Wikipedia, *Honours Bachelor thesis, The University of Sydney Australia* (2008).

[71] R. Al-Rfou, V. Kulkarni, B. Perozzi and S. Skiena, Polyglot-NER: Massive multilingual named entity recognition, in: *Proceedings of the 2015 SIAM International Conference on Data Mining*, SIAM, 2015, pp. 586–594.

[72] A. Ghaddar and P. Langlais, Winer: A wikipedia annotated corpus for named entity recognition, in: *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 413–422.

[73] A. Ghaddar and P. Langlais, Transforming Wikipedia into a Large-Scale Fine-Grained Entity Type Corpus, in: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[74] K. Kettunen and T. Ruokolainen, Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection, in: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, ACM, 2017, pp. 181–186.

[75] C.-T. Tsai, S. Mayhew and D. Roth, Cross-lingual named entity recognition via Wikification, in: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 219–228.

[76] J. Nothman, N. Ringland, W. Radford, T. Murphy and J.R. Curran, Learning multilingual named entity recognition from Wikipedia, *Artificial Intelligence* **194** (2013), 151–175.

[77] J. Hoffart, F.M. Suchanek, K. Berberich and G. Weikum, YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, *Artificial Intelligence* **194** (2013), 28–61.

[78] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter and R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[79] J. Pennington, R. Socher and C. Manning, GloVe: Global Vectors for Word Representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162. https://www.aclweb.org/anthology/D14-1162.

[80] J.L. Fleiss, Measuring nominal scale agreement among many raters., *Psychological Bulletin* **76** (1971), 378–382.

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