Who is Mona L.? Identifying Mentions of Artworks in Historical Archives

Nitisha Jain and Ralf Krestel

Hasso Plattner Institute, University of Potsdam, Germany
firstname.lastname@hpi.de

Abstract. Named entity recognition (NER) plays an important role in many information retrieval tasks, including automatic knowledge graph construction. 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 art historical archives, the recognition of titles of artworks as named entities is of high importance. In this work, we focus on identifying mentions of artworks, e.g. paintings and sculptures, from historical archives. Current state of the art NER tools are unable to adequately identify artwork titles due to the particular difficulties presented by this domain. The scarcity of training data for NER for cultural heritage poses further hindrances. To mitigate this, we propose a semi-supervised approach to create high-quality training data by leveraging existing cultural heritage resources. Our experimental evaluation shows significant improvement in NER performance for artwork titles as compared to baseline approach.

Keywords: named entity recognition · cultural heritage data

1 Artwork Mentions in Historical Archives

Named entity recognition (NER) is a key component for information extraction pipelines that aims to identify the named entities in text and classify them into pre-defined categories. NER serves as an important step for various semantic tasks, such as knowledge base creation, text based search, relation extraction and question answering, among many others. There is a large body of existing work on improving its performance, with the recent approaches based on machine learning techniques. However, most efforts have focused only on some common categories of named entities, i.e., person, organization, location, and date. Moreover, state of the art NER systems are trained on a few well-established corpora available for the task such as the CoNNL datasets [8] or OntoNotes [5]. Although these systems attain good results for generic tasks, their performance and utility is essentially limited due to the specific training. Thus, it comes as no surprise that it has been a challenge to adapt NER systems for identifying domain-specific named entity categories with reasonable accuracy [6].

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 cultural data with the principles of linked open data[1] gaining momentum in the cultural heritage domain [11]. 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. Efforts have been made to digitize historical as well as recent art related texts such as auction catalogues, art books and exhibition catalogues [3]. In such resources, cultural objects, mainly artworks, are often described with help of unstructured text narratives. The identification and extraction of the mentions of artworks from such text descriptions can serve various important use cases, such as facilitate search and browsing in digital resources, help art historians with tracking of provenance of artworks and enable wider semantic text exploration for digital cultural resources.

In this paper, we refer to the named entities depicting the titles of artworks to be of type *title*. These titles could have been assigned by artists, by collectors, art historians, or other domain experts. Due to the ambiguities that are inherent in artwork titles, their identification from texts is a challenging task. As an example, consider the painting titled '*Girl before a mirror*' by famous artist Pablo Picasso. This title merely describes in an abstract manner what is being depicted in the painting and thus, it is hard to identify it as a named entity without knowing the context of its mention. Yet, such descriptive titles are common in the art domain, as are abstract titles such as *'untitled'*. In this work, we focus on identifying mentions of artworks from unstructured text in art historical archives. Due to the innate complexity of this task, NER models need to be trained with domain-specific named entity annotations. As such, the unavailability of high-quality training data for the cultural heritage domain is one of the biggest hindrances for this task. We address this gap by proposing techniques for generating annotations for NER via a semi-automated approach from a large corpus of art related documents, while leveraging existing art resources that are integrated in popular knowledge bases, such as Wikidata [12].

2 Named Entity Recognition for Artworks

Identification of mentions of artworks seems, at first glance, to be no more difficult than detecting mentions of persons or locations. But the special characteristics of artwork titles makes this a complicated task which requires significant domain expertise. This section illustrates three types of errors that arise when trying to recognize artwork mentions in practice.

[1] Linked Open Data: http://www.w3.org/DesignIssues/LinkedData
[2] OpenGLAM: http://openglam.org
[3] Europeana: http://europeana.eu

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

Incorrect Named Entity Boundary Detection. Often, artworks have long and descriptive titles, e.g., a painting by Van Gogh titled *'Head of a peasant woman with dark cap'*. If this title is mentioned in text without any formatting indicators, it is likely that the boundaries may be wrongly identified and the named entity be tagged as *'Head of a peasant woman'*, which is also the title of a different painting by Van Gogh. In fact, Van Gogh had created several paintings with this title in different years. For such titles, it is common that location or time indicators are appended to the titles (by the collectors or curators of museums) in order to differentiate the artworks. However, such indicators are not a part of the original title and should not be included within the scope of the named entity.

Incorrect Named Entity Type Tagging. 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 portraits and self-portraits. 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. Numerous old paintings are portraits of the prominent personalities of those times and are named after them such as *'King George III'*, *'Queen Anne'* and so on — 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 type *location*.

3 Related Work

In the absence of manually curated NER annotations, the adaptation of existing NER solutions to the art and cultural heritage domain faces multiple challenges, some of them being unique to this domain. Seth et al. [10] 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. In [7], Rodriquez et al. 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. On similar lines, Ehrmann et al. [4] offer a diachronic evaluation of various NER tools for digitized archives of Swiss newspapers. However, none of the existing works have focused on the task of identifying artwork titles that are highly relevant as a named entity type for the art domain. Moreover, previous works have merely compared the performance of existing NER systems, whereas in this work, we aim to improve the performance of NER systems for cultural heritage with the help of domain-specific high-quality training data.

Although there is increasing effort to publish cultural heritage collections as linked data [2, 9, 3], to the best of our knowledge, there is no annotated dataset available for facilitating NER in this domain yet. This work proposes techniques to generate a high-quality training corpus in a scalable and semi-supervised manner and demonstrates that NER systems can be trained to identify mentions of artworks with notable performance gains.

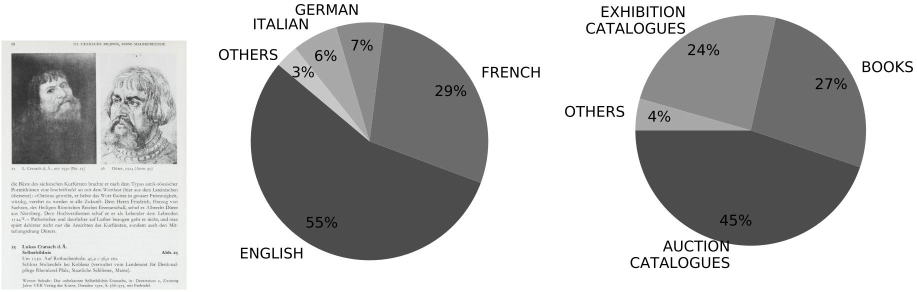
4 Annotating Complex Named Entity Types

NER Model. None of the existing NER systems can identify titles of artworks as named entities out of the box. The closest NER category to artwork titles was found in the SpaCy[4] library as *work_of_art*, which refers not only to artworks such as paintings and sculptures, but also covers a large variety of others cultural heritage objects such as movies, plays, books, songs etc. Although the pre-trained SpaCy model performed poorly for cultural heritage domain, we have used this as a naive baseline for the lack of better alternatives. In order to improve the identification of named entities of type *title*, training on high-quality annotated training datasets is imperative and for this purpose, the baseline SpaCy NER model was leveraged for domain-specific re-training. Due to the steep costs and efforts of manual annotations, we aimed to generate a large corpus of annotated data in a semi-automated fashion from our dataset. It is to be noted that the proposed techniques for improving the quality of NER training data are independent of the NER model used for the evaluation. Thus, SpaCy can be substituted with any other re-trainable NER system.

Training Dataset. The underlying dataset for this work is a large collection of art historical documents that have been recently digitized. The collection consists of different types of documents — auction catalogues, full texts of art books related to particular artists or art genres, catalogues of art exhibitions and other documents. A sample document[5] is shown in Fig.1a. The auction and exhibition catalogues contain semi-structured and unstructured texts that describe artworks on display, mainly paintings and sculptures. Art books may contain more unstructured text about the origins of artworks and their creators. Fig. 1c shows the distribution of the different types of documents in the dataset. The pages of these catalogues and books were scanned with OCR and each page was converted to an entry stored within a search index. Due to the limitations of OCR, the dataset suffers from noise and does not retain its rich original formatting information which would have been quite useful for analysis. The

[4] SpaCy: https://spacy.io/, version 2.1.3

[5] from the exhibition catalogue “Lukas Cranach: Gemälde, Zeichnungen, Druckgraphik” (https://digi.ub.uni-heidelberg.de/diglit/koepplin1974bd1/0084/image)



78 III. CRANACHS BILDNIS, SEINE MALERFREUNDE

25 L. Cranach d. Ä., um 1550 (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[78].» 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.

(a) Sample Document

(b) Languages

(c) Document types

Fig. 1: Dataset Characteristics

dataset consists of texts in a number of different languages, which adds additional complexity to the NER task. English, French, German and Italian account for the majority of the languages as shown in Fig. 1b, while Dutch, Spanish, Swedish and Danish were also recognized in a sizeable number of entries. In this work, however, we avoid the multi-lingual analysis for the sake of simplicity and focus on the NER task for English documents. After initial pre-processing including the removal of non alpha-numeric characters, the dataset consisted of a total of 117,912 entries in English, which was then transformed into annotated NER data.

Named Entity Annotations with High Precision. In order to match and correctly tag the artwork titles present in our dataset as named entities of type *title*, we leveraged cultural resources that have been integrated into popular knowledge bases. As a first step, available resources from Wikidata were collected to generate a large entity dictionary or *gazetteer* of titles of artworks. Integrating other sources, such as art-related ontologies or lists from museums is also possible. To generate the entity dictionary for titles, Wikidata was queried with the Wikidata Query Service[6] for names of artworks, specifically for names of paintings and sculptures. In order to match the original non-English titles of artworks, titles belonging to other major languages present in our dataset were also added. Many of the titles were highly generic, for instance, '*Italian*', '*Winter*', '*Landscape*' etc., therefore, the titles consisting of only one word filtered out. Since quite a few artwork titles were identical to location names that could lead to incorrect name entity type tagging, such titles were also ignored. A combined list of approximately 15,000 titles in different languages were obtained, with the majority of them being in English.

Named Entity Annotations with High Recall. As discussed in Section 2, partial matching of artwork titles can lead to ambiguities. Due to the limitations

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

of the naive NER model there were several 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 example, from the text *"..the subject of the former is not Christ before Caiaphas, as stated by Birke and Kertész, 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 *title* entities after the boundary corrections. Through this technique, a number of missed mentions of artwork titles were added to the training dataset, thus improving the recall of the annotations and in turn, influencing NER performance positively.

5 Evaluation

Experimental Setup. In order to evaluate the impact on NER performance with improvements in quality of the training data, we trained the baseline NER model for the new entity type *title* on different variants of training data:
High-precision : Annotations obtained by matching Wikidata titles.
High-recall : Additional annotations from named entity boundary corrections.

The number of annotations (training set size) for each of the datasets are shown in Table 1. An NER model was obtained by training with the above datasets for 10 epochs, with the training data batched and shuffled before every iteration. The performance of the trained NER models was compared with the *Baseline* NER model i.e. the pre-trained SpaCy model without any specific annotations for artwork titles. In the absence of a gold standard dataset for NER for artwork titles, we performed manual annotations to obtain a test dataset for evaluation.

Manual Annotations for Test Dataset. For generating a test dataset, a set of texts were chosen at random from the dataset, while making sure that this text was representative of the different types of documents in our corpus. This test data consisted of 544 entries (with one or more sentences per entry) and was carefully excluded from the training dataset. The titles of paintings and sculptures mentioned in this data were then manually identified and tagged as named entities of type *title*. The annotations were performed by two non-expert annotators independently of each other in 3–4 person hours with the help of Enno[7] tool. The inter-annotator agreement in terms of the Fleis-kappa and Krippendorf-kappa scores were calculated to be -1.86 and 0.61 respectively. The poor inter-annotator agreement reflected by these scores reaffirmed that the task of annotating the artwork titles is difficult, even for humans. In order to obtain the gold standard test dataset for the evaluation of NER models, the disagreements were manually sorted out with the help of web search, resulting in a total of 144 entities being positively tagged as *title*.

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

Table 1: Performance of NER Models Trained on Different Annotated Datasets

Train Dataset	Size	*Strict*			*Relaxed*		
		P	R	F1	P	R	F1
Baseline	–	.14	.06	.08	.22	.08	.12
High-precision	226,801	.20	.12	.15	.32	.20	.25
High-precision + High-recall	413,932	**.23**	**.22**	**.23**	**.39**	**.41**	**.40**

Evaluation Metrics. 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 [8], where the entities' boundaries were matched exactly. The MUC NER task [1] 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 entities of type *title* and therefore, it was sufficient to check only for the boundary matches of the identified entities. 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 metrics allowed for comparison of the entities despite errors due to wrong chunking of the named entities in the text (Section 2).

Results and Discussion. The results shown in Table 1 demonstrate significant improvement in performance for the NER models that were re-trained with relevant 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. With the benefit of domain-specific and entity-specific annotations generated from the Wikidata entity dictionaries, the high-precision NER model was able to correctly identify many artwork titles. The performance was further boosted after including the high-recall dataset having additional annotations obtained with the help of boundary corrections. This illustrates the importance of quality of the NER training data for challenging domains. Our approach to generate high-quality annotations in semi-automated manner from a domain-specific corpus is an important contribution towards this direction.

6 Conclusion

In this work we proposed an approach to identify artwork mentions from art historic archives. We motivated the need for NER training on high-quality annotations and proposed techniques for generating the relevant training data for this task in 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 historical archives, supervised NER models can be trained to perform well. Furthermore, our approach is not limited to the cultural heritage domain but can be adapted for other domain-specific NER tasks, where there is also shortage of annotated training data. As future work we would like to apply our techniques for named entity recognition to other important entities and perform entity-centric text exploration for cultural heritage resources.

Acknowledgements. We thank the Wildenstein Plattner Institute[8] for providing the corpus used in this work.

References

1. Chinchor, N.: Overview of MUC-7. In: Proceedings of the Seventh Message Understanding Conference (MUC-7) (1998)
2. De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., Schreiber, G.: Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In: Extended Semantic Web Conference. pp. 733–747. Springer (2012)
3. Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., ter Weele, W., Wielemaker, J.: The Rijksmuseum Collection as Linked Data. Semantic Web **9**(2), 221–230 (2018)
4. Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic evaluation of ner systems on old newspapers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). pp. 97–107 (2016)
5. Pradhan, S., Moschitti, A., Xue, N., Ng, H.T., Björkelund, A., Uryupina, O., Zhang, Y., Zhong, Z.: Towards Robust Linguistic Analysis using OntoNotes. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. pp. 143–152 (2013)
6. Prokofyev, R., Demartini, G., Cudré-Mauroux, P.: Effective Named Entity Recognition for Idiosyncratic Web Collections. In: Proceedings of the 23rd international conference on World Wide Web (WWW). pp. 397–408. ACM (2014)
7. Rodriquez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of Named Entity Recognition tools for raw OCR text. In: Konvens. pp. 410–414 (2012)
8. Sang, E.F.T.K., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Development **922**, 1341 (1837)
9. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In: Extended Semantic Web Conf. pp. 593–607. Springer (2013)
10. Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. Digital Scholarship in the Humanities **30**(2), 262–279 (2013)
11. Van Hooland, S., Verborgh, R.: Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish your Metadata. Facet publishing (2014)
12. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledge base. Commun. ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489

[8] https://wpi.art