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

Nitisha Jain and Ralf Krestel

Hasso-Plattner-Institut, Prof.-Dr.-Helmert-Str. 23, 14482 Potsdam, Germany
firstname.lastname@hpi.de

Abstract. Named entity recognition (NER) plays an important role in many natural language processing tasks, including automatic knowledge graph construction and ontology generation. 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 digitized versions of art 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 from knowledge bases such as Wikidata. Our experimental evaluation shows significant improvement in NER performance for artwork titles as compared to baseline approaches.

Keywords: named entity recognition · semi-supervised learning · cultural heritage data · knowledgebase creation

1 Artwork Mentions in Historical Archives

Named entity recognition (NER) is a key component for many 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 — classically, with the help of statistical methods and recently, with the help of machine learning approaches. 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 [26, 25] or OntoNotes [22]. 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 [21, 23].

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 [33]. 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 [10, 5]. 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 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 or, in the case of certain old and ambiguous artworks, by collectors, art historians, or other domain experts. Due to the ambiguities that are inherent in artwork titles, their identification from texts is a challenging task. As an example, consider the painting titled '*Girl before a mirror*' by famous artist Pablo Picasso. This title merely describes in an abstract manner what is being depicted in the painting and thus, it is hard to identify it as a named entity without knowing the context of its mention. Similarly, consider the painting with the title '*Head of a woman*' — such phrases can be hard to be distinguished as named entities from the surrounding text due to their generality. Yet, such descriptive titles are common in the art domain, as are abstract titles such as *'untitled'*.

To circumvent ambiguities present in art-related documents for human readers, artwork titles are typically formatted in special ways : they are distinctly highlighted with capitalization, quotes, italics or boldface fonts, etc. which provide the required contextual hints to identify them as titles. However, the presence of these formatting cues cannot be assumed or guaranteed, especially in texts from art historical archives, due to adverse effects of scanning errors on the quality of digitized resources [11]. 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.

In this work, we focus on identifying the 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, such that the models can learn important textual features to achieve the desired

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

results. As such, the unavailability of high-quality training data for the cultural heritage domain is one of the biggest hindrances for this task. We address this gap by proposing techniques for generating annotations for NER via a semi-automated approach from a large corpus of art related documents. To this end, we leverage existing art resources, namely titles of artworks, that are integrated in popular knowledge bases, such as Wikidata [34]. 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. Experimental evaluations demonstrate substantial improvement in NER performance (more than doubling the F1 score) when trained with the high-quality annotations generated through our methods.

2 Named Entity Recognition for Works of Art

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 titles of artworks makes this a complicated task which requires significant domain expertise to tackle. We want to illustrate the difficulties that arise when trying to recognize artwork mentions in practice. There are three types of errors that can be distinguished — Failure of detection of a *title* named entity, incorrect detection of the named entity boundaries, and incorrect tagging of the *title* named entity with a wrong type.

2.1 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.

2.2 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. On the other hand, for the

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

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

From these examples it is apparent that NER for artwork titles is a non-trivial task. We discuss previous efforts related to this problem in the next section.

3 Related Work

Named entity recognition, being important for many NLP tasks, has been the subject of numerous research efforts. Several prominent NER systems have been developed that have achieved near human performance for the few most common entity types on certain datasets. Previously, the best performing NER systems were trained through feature-engineered supervised techniques such as Hidden Markov Models (HMM), Support Vector Machines (SVM) and Conditional Random Fields (CRF) [36, 19, 18, 1]. In the past decade, such systems have been bested by unsupervised 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 [2, 9, 28], and Convolutional Neural Networks (CNN) for character level representation [13, 16, 6] 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 feature engineering approaches. These models have achieved state of the art results on multilingual CoNNL 2002 and 2003 datasets [17, 35]. 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. In most cases, these systems fail to adapt well to new domains and different named entity categories [23, 21].

There have been previous efforts towards domain specific NER as well, mainly in the biomedical domain. NER systems have been used to identify the names of drugs, proteins and genes [12, 31, 15]. But since these NER techniques rely on specific resources such as carefully curated lists for drug names [14] or biology and microbiology NER datasets [7, 4], they are highly specific solutions geared towards biomedical domain.

For the NER systems that have been developed until now, the primary research focus has been on the improvement of the NER model architecture 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 popular available benchmark datasets. However, this approach is not feasible for domain-specific tasks, such as for the identification of artwork titles in cultural heritage domain, since the generation of large domain-specific manually annotated data is expensive in terms of human labour while also requiring significant domain expertise.

In the absence of manually curated NER annotations, the adaptation of existing NER solutions to the art and cultural heritage domain faces many challenges, some of them being unique to this domain. Seth et al. [32] 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. [27] 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. [24] 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. However, none of the existing works have focused on the task of identifying artwork titles which are one of the most important named entities for the art domain. Moreover, previous works have merely compared the performance of existing NER systems, whereas in this work, we aim to improve the performance of NER systems for cultural heritage with the help of domain-specific high-quality training data.

Although there is increasing effort to publish cultural heritage collections as linked data [3, 29, 5], to the best of our knowledge, there is no annotated dataset available for NER in this domain. This work proposes novel 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

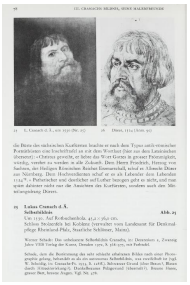
In this section we discuss our approach for generating high-quality training data for the NER task without the need for manual annotations. These techniques were geared towards tackling the challenges presented by noisy corpora that are typical of art historical archives, although they can be applicable for other domains as well.

4.1 NER Model

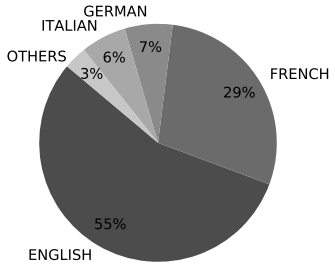
None of the existing NER systems can identify titles of artworks as named entities out of the box. The closest NER category to artwork titles was found in the SpaCy[4] library as *work_of_art*. This category refers not only to artworks such as paintings and sculptures, but also covers a large variety of cultural heritage objects including movies, plays, books, songs etc. For the lack of alternatives, we have leveraged this NER category in our work for setting up a naive baseline with which we compare the improvements in NER performance.

The SpaCy library for natural language processing was employed for tokenization and chunking of the texts before the identification of the named entities. The pre-trained English model of SpaCy has been trained on Ontonotes5 dataset[5] which consists of different types of texts including telephone conversations, newswire, newsgroups, broadcast news etc. Since this dataset is considerably different from historical art document collections, the pre-trained NER model showed poor performance for named entity recognition in the cultural heritage domain, even for the common named entity types (*person*, *location* and *organization*). With regards to artwork titles, very few were identified as named entities and many among those were wrongly tagged as names of persons or locations, instead of being correctly categorized as *work_of_art*. The pre-trained SpaCy NER model will be referred to as the baseline model. In order to improve the identification of named entities of type *title*, training on high-quality annotated training datasets is imperative and for this purpose, the baseline NER model was leveraged for re-training. Due to the steep costs and efforts of human annotations, we aimed to generate a large corpus of annotated data in a semi-automated fashion from our dataset. It is to be noted that the techniques for improving the quality of NER training data that are proposed in this work are independent of the NER model used for the evaluation. Thus, SpaCy is merely a tool which can be substituted with any other re-trainable NER system.

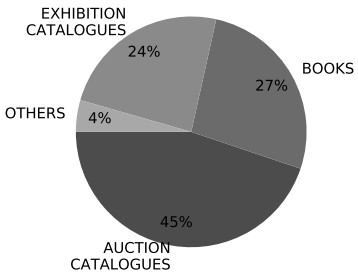
[4] SpaCy: https://spacy.io/, version 2.1.3
[5] https://catalog.ldc.upenn.edu/LDC2013T19



(a) Sample Document



(b) Languages



(c) Document types

Fig. 1: Dataset Characteristics

4.2 Training Dataset

The underlying dataset for this work is a large collection of art historical documents that have been recently digitized. A sample document[6] is shown in Fig. 1a. The collection consists of different types of documents: auction catalogues, full texts of art books related to particular artists or art genres, catalogues of art exhibitions and other documents. The auction and exhibition catalogues contain semi-structured and unstructured texts that describe artworks on display, mainly paintings and sculptures. Art books may contain more unstructured text about the origins of artworks and their creators. Fig. 1c shows the proportion of the different kinds of documents in the dataset. The pages of these catalogues and books were scanned with OCR and each page was converted to an entry stored within an elastic search index. Due to the limitations of OCR, the dataset did not retain its rich original formatting information which would have been very useful for analysis. In fact, the data suffers from many spelling and formatting mistakes that need to be appropriately handled.

The dataset consists of texts in more than 30 different languages, which adds additional complexity to the NER task. English, French, German and Italian account for 97% of the languages as shown in Fig. 1b. 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 unrecognizable characters, the dataset consisted of a total of 117,912 entries in English and stored in multiple text files each having around 500 entries. This sizeable dataset was transformed into annotated NER data as described in following sections.

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

4.3 Named Entity Annotations with High Precision

In order to match and correctly tag the titles present in our corpus as named entities of type *title*, we leveraged cultural resources that have been integrated into the popular knowledge bases. As a first step, we collected available resources from Wikidata to generate a large entity dictionary or *gazetteer* of titles of artworks in an automatic way. Integrating other sources, such as art-related ontologies or lists from museum resources is also possible. To generate the entity dictionary for titles, Wikidata was queried with the Wikidata Query Service[7] for names of artworks, specifically for names of paintings and sculptures. Since our 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 the longest title having 221 characters in total (*'Predella Panel Representing the Legend of St. Stephen ... '*). Many of the titles were highly generic, for instance, '*Italian*', *'Winter'*, *'Landscape'* etc., therefore, we filtered out the titles having only one word from the list. Since many artwork titles are identical to location names which can lead to many errors while tagging the named entity to the correct type, such titles were also ignored. The large variety and ambiguity observed in the titles extracted from Wikidata further confirmed that the NER for artwork titles is a non-trivial task. A combined list of approximately 15,000 titles in different languages were obtained, majority of them being in English. Due to inconsistencies in the capitalization of the words in the title found on Wikidata, as well as in the mention of titles in our dataset, the titles had to be uniformly lower-cased to enable matching. To circumvent the issue of false positives in annotations, the entire dataset was not searched for matches with the *title* entity dictionary. Firstly, the named entities of all categories, as identified by SpaCy NER model, were extracted from the dataset. Thereafter, the successful matches for the extracted entities were tagged as *title* named entity types. Even though some named entities were inadvertently missed with this approach, it facilitated the generation of high-precision annotations from the underlying dataset from which the NER model could learn useful features.

4.4 Named Entity Annotations with High Recall

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

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

and extend the boundaries of the named entities to obtain the complete and correct titles. For a given text, a separate list of matches with the artwork titles in entity dictionary over the entire text was maintained as *spans* (starting and ending character offsets), in addition to the extracted named entities. It is to be noted that the list of *spans* included many false positives due to matching of generic words and phrases that were not named entities. The overlaps between the two lists were considered, if a *span* was a super-set of a named entity, the boundary of the identified named entity was extended as per the *span* offsets. For example, from the text *"..The subject of the former (inv. 3297) is not Christ before Caiaphas, as stated by Birke and Kertsz, 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, many 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.

4.5 Wikipedia-Derived Silver Standard Annotations

Despite efforts for high accuracy, one of the major limitations of generating named entity annotations from art historical archives is the presence of errors in the training data. To enable an NER model to learn the textual indicators present in the dataset for identification of titles, we further augmented our training dataset with clean and well-structured silver standard[8] annotations derived from Wikipedia articles that proved very useful for NER training. Several previous works have utilized the anchor texts and the tagged categories present in Wikipedia articles to transform sentences into named entity annotations [30, 20, 8]. However, in this work, we followed a different approach, where we mined relevant sentences from different Wikipedia articles that, in turn, referred to a Wikipedia article on an artwork.

To find such sentences, firstly, we identified all the artwork titles having a corresponding Wikipedia page in English; a total of 2808 pages were found. For each Wikipedia page referring an artwork, the back-links, i.e. the URLs of the pages that referred to this page were collected. The pages were searched for the relevant sentences that contained an outgoing link to the Wikipedia page of the artwork, while also making sure that anchor text of the outgoing link was identical to the title of the artwork. These sentences were extracted and the anchor texts of the sentences was tagged as a *title* named entity, serving as accurate annotations for this category. Through this process, a total of 1099 sentences were added as silver standard annotation data to the training set. This data provided correct and precise textual patterns that were highly indicative of the artwork titles and led to a further boost in NER performance. We discuss

[8] 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.

the experimental evaluation and performance gains for the different variants of the training data in the next section.

5 Experimental Evaluation

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

5.1 Experimental Setup

In order to evaluate and compare the impact on NER performance with improvements in quality of the training data, we trained the baseline NER model for the new entity type *title* on different variants of training data as follows:
High-precision: Training dataset with annotations obtained with matching of Wikidata titles,
High-recall: Training dataset with additional annotations from named entity boundary corrections,
Wikipedia-derived: Training dataset augmented with silver standard annotations derived from Wikipedia.
The number of annotations present in the above training datasets (*Size*) is shown in Table 1. An NER model was obtained by training with each of the above datasets for 10 epochs, with the training data batched and shuffled before every iteration. The performance of the trained NER models was compared with the *Baseline* NER model (which was pre-trained without any specific annotations for artwork titles). Since the named entity type *title* was not applicable for the baseline model, a match with the entity category *work_of_art* was considered as a true positive. In the absence of a gold standard dataset for NER for artwork titles, we performed manual annotations and generated a test dataset on which the models could be suitably evaluated.

5.2 Manual Annotations for Test Dataset

For generating a test dataset, a set of texts were chosen at random from the dataset, while making sure that this text was representative of the different types of document collections in the overall corpus. This test data consisted of 544 entries (with one or more sentences per entry) and was carefully excluded from the training dataset. The titles of paintings and sculptures mentioned in this data were then manually identified and tagged as named entities of 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[9] tool and their respective annotations were compared afterwards. The task of manual annotation was found challenging due to the inherent ambiguities in the dataset (Section 2)

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

and lack of domain expertise. The annotators disagreed on the tagging of certain phrases as titles on multiple occasions. The inter-annotator agreement in terms of the Fleis-kappa and Krippendorf-kappa scores were calculated to be -1.86 and 0.61 respectively. (A negative Fleis-kappa score indicates poor agreement, while Krippendorf-kappa values for data should be above 0.667 to be considered useful.) The poor inter-annotator agreement reflected by these scores 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 and a total of 144 entities were positively tagged as *title*.

5.3 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[10], where the entities' boundaries were matched exactly. The MUC NER task[11] 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. Since there are many ambiguities involved with entity boundaries of artwork titles, as discussed in Section 2.2, we evaluated the NER models with both strict 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, and F1 scores obtained for the NER models trained with different training dataset variants are shown in Table 1.

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
High-precision + High-recall + Wikipedia-derived	415,031	**.26**	**.25**	**.26**	**.43**	**.42**	**.43**

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

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

5.4 Results and Discussion

The results demonstrated definitive improvement in performance for the NER models that were trained with annotated data as compared to the baseline performance, as expected. 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 the training on high-recall dataset having additional annotations obtained with the help of boundary corrections. More importantly, it is encouraging to see that with the addition of small corpus of silver standard annotation data derived from Wikipedia, the NER model was able to achieve substantially better results. This illustrates the positive impact of the quality of the NER training data for challenging domains and motivates the importance of training on high-quality datasets. Our approach to generate such high-quality annotations in semi-automated manner from a domain-specific corpus is an important contribution towards this direction. Moreover, the remarkable improvement for NER performance achieved for a novel and challenging named entity of type *title*, proves the effectiveness of our approach.

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. It would be interesting to leverage named entities to mine interesting patterns about artworks and artists, which may facilitate the creation of a comprehensive knowledge base for this domain.

References

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

2. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(Aug), 2493–2537 (2011)
3. 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)
4. Delėger, L., Bossy, R., Chaix, E., Ba, M., Ferrė, A., Bessieres, P., Nėdellec, C.: Overview of the bacteria biotope task at bionlp shared task 2016. In: Proceedings of the 4th BioNLP shared task workshop. pp. 12–22 (2016)
5. 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)
6. Gillick, D., Brunk, C., Vinyals, O., Subramanya, A.: Multilingual language processing from bytes. arXiv preprint arXiv:1512.00103 (2015)
7. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtIvE: critical assessment of information extraction for biology (2005)
8. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence **194**, 28–61 (2013)
9. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
10. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Kettula, S.: MuseumFinland : Finnish Museums on the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web pp. 224–241 (2005)
11. Kettunen, K., Ruokolainen, T.: Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. In: Proc.of the 2nd Intl. Conference on Digital Access to Textual Cultural Heritage. pp. 181–186. ACM (2017)
12. Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: 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. pp. 70–75. Citeseer (2004)
13. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
14. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al.: DrugBank 3.0: A comprehensive resource for omics research on drugs. Nucleic acids research **39**(suppl_1), D1035–D1041 (2010)
15. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., et al.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics **7**(1), S2 (2015)
16. Kuru, O., Can, O.A., Yuret, D.: Charner: Character-level named entity recognition. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 911–921 (2016)
17. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
18. Li, Y., Bontcheva, K., Cunningham, H.: SVM based learning system for information extraction. In: International Workshop on Deterministic and Statistical Methods in Machine Learning. pp. 319–339. Springer (2004)
19. Malouf, R.: Markov models for language-independent named entity recognition. In: Proc. of the 6th Conference on Natural Language Learning (CoNLL) (2002)

20. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence **194**, 151–175 (2013)
21. Poibeau, T., Kosseim, L.: Proper name extraction from non-journalistic texts. Language and computers **37**, 144–157 (2001)
22. 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)
23. 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)
24. 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)
25. Sang, E.F.T.K., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Development **922**, 1341 (1837)
26. Sang, T.K., Erik, F.: Introduction to the CoNLL-2002 sShared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002/Roth, Dan [edit.]. pp. 155–158 (2002)
27. Segers, R., Van Erp, M., Van Der Meij, L., Aroyo, L., Schreiber, G., Wielinga, B., van Ossenbruggen, J., Oomen, J., Jacobs, G.: Hacking history: Automatic historical event extraction for enriching cultural heritage multimedia collections. In: Proc. of the 6th Intl. Conference on Knowledge Capture (K-CAP). pp. 26–29 (2011)
28. Shao, Y., Hardmeier, C., Nivre, J.: Multilingual named entity recognition using hybrid neural networks. In: The Sixth Swedish Language Technology Conference (SLTC) (2016)
29. 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)
30. Tsai, C.T., Mayhew, S., Roth, D.: Cross-lingual named entity recognition via Wikification. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. pp. 219–228 (2016)
31. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association **14**(5), 550–563 (2007)
32. 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)
33. Van Hooland, S., Verborgh, R.: Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet publishing (2014)
34. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489
35. Yadav, V., Sharp, R., Bethard, S.: Deep affix features improve neural named entity recognizers. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 167–172 (2018)
36. Zhou, G., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 473–480. Association for Computational Linguistics (2002)