Are End-to-End Systems Really Necessary for NER on Handwritten Document Images?

Oliver Tüselmann (✉)[0000−0002−8892−3306], Fabian Wolf[0000−0001−8842−3718], and Gernot A. Fink[0000−0002−7446−7813]

Department of Computer Science, TU Dortmund University,
44227 Dortmund, Germany
{oliver.tueselmann, fabian.wolf, gernot.fink}@cs.tu-dortmund.de

Abstract. Named entities (NEs) are fundamental in the extraction of information from text. The recognition and classification of these entities into predefined categories is called Named Entity Recognition (NER) and plays a major role in Natural Language Processing. However, only a few works consider this task with respect to the document image domain. The approaches are either based on a two-stage or end-to-end architecture. A two-stage approach transforms the document image into a textual representation and determines the NEs using a textual NER. The end-to-end approach, on the other hand, avoids the explicit recognition step at text level and determines the NEs directly on image level. Current approaches that try to tackle the task of NER on segmented word images use end-to-end architectures. This is motivated by the assumption that handwriting recognition is too erroneous to allow for an effective application of textual NLP methods. In this work, we present a two-stage approach and compare it against state-of-the-art end-to-end approaches. Due to the lack of datasets and evaluation protocols, such a comparison is currently difficult. Therefore, we manually annotated the known IAM and George Washington datasets with NE labels and publish them along with optimized splits and an evaluation protocol. Our experiments show, contrary to the common belief, that a two-stage model can achieve higher scores on all tested datasets.

Keywords: Named entity recognition · Document image analysis · Information retrieval · Handwritten documents

1 Introduction

Named entities (NEs) are objects in the real world, such as persons, places, organizations and products, that can be referred to by a proper name. They are known to play a fundamental role in the extraction of information from text. Thereby, NEs are not only used as a first step in question answering, search engines or topic modeling, but they are also one of the most important information for indexing documents in digital libraries [31]. The extraction of this information from text is called Named Entity Recognition (NER) and is an important field of research in Natural Language Processing (NLP). Over the past

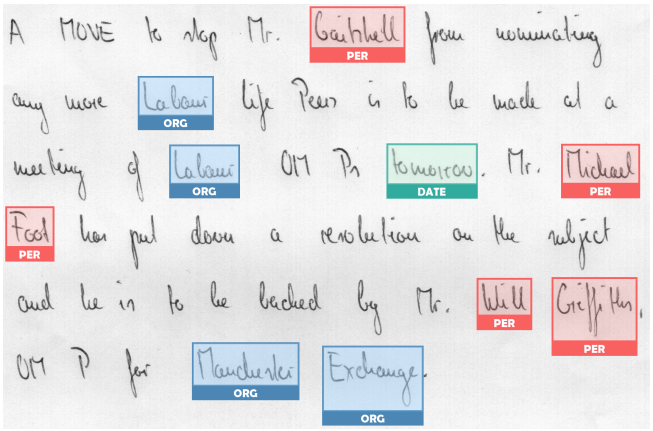

Fig. 1: An example for named entities on a handwritten document image from the IAM database.

few years, impressive progress has been made in this area [32]. The Document Image Analysis community, on the other hand, focuses on the recognition and retrieval of handwritten document images and less on the extraction of information from those. Therefore, there are only a few publications that try to tackle NER on document images. Figure 1 shows an example for NEs on a handwritten document image.

An intuitive approach for NER on document images is to combine the advances from the Document Image Analysis and NLP domain, using a two-stage model. The first stage is the transformation of the document into a textual representation. After that, the relevant NEs are extracted using an NLP model. Unfortunately, despite advances in machine learning, the recognition approaches are still not perfect and could produce high Character Error Rates (CERs) and Word Error Rates (WERs), especially on handwritten documents. An interesting question is whether and how errors from the recognition stage affect the performance of an NLP model. Especially in the last few years, there have been several publications that have studied the effect of Optical Character Recognition (OCR) errors on NLP tasks [6,13,14]. It is generally agreed that OCR errors have a negative impact on the performance of NLP models and that their performance degrades when CERs and WERs increase. To overcome the problem of error propagation, an end-to-end architecture is often used. Such an approach avoids the explicit recognition step and predicts the NEs directly from word images. Even though this approach can solve the error propagation problem, the architecture has the fundamental drawback of not using the most powerful advantages from the NLP domain. These are the use of pre-trained word embeddings on large datasets and transfer learning [32]. A word embedding is an encoded knowledge base containing semantic relations between words. Transfer learning refers to the process of applying knowledge from one task to another.

Mainly due to the small amount of digitized handwritten documents and the variability of handwriting, it is not yet possible to apply the performance of word embeddings in the word image domain. Also, the use of synthetically generated documents cannot solve the lack of training data, since transfer learning is still a challenging task in the handwriting domain [12]. There are already works dealing with the mapping of word images to pre-trained textual word embeddings [16,29]. Nevertheless, they only work with static word embeddings, being limited. Therefore, end-to-end approaches can only learn from the available training data in the image domain and take into account either extremely limited or no prior semantic knowledge about words.

Even though end-to-end as well as two-stage approaches have their advantages and disadvantages for NER on word-segmented handwritten document images containing unstructured text, only end-to-end architectures are published so far. This is motivated by the assumption that handwriting recognition is too erroneous to allow for an effective application of textual NLP methods. However, Hamdi et al. successfully apply a two-stage approach in [14] and showed recently on a machine-printed Finnish historical newspaper with a CER of 6.96% and a WER of 16.67% that there is only a marginal decrease in the F1-score compared to the original text (89.77% to 87.40%). Since such error rates can also be achieved with a state-of-the-art recognizer on most handwritten datasets, we investigate in this work a two-stage approach and compare it to end-to-end methods from the literature. Such a comparison is not straightforward as there are almost no established datasets and evaluation protocols. Most approaches from the literature use their own, unpublished annotations and evaluation protocols, making a direct comparison impossible. In order to provide comparability, we also present and publish NE annotations for the well-known George Washington (GW) and IAM datasets as well as an evaluation protocol [1].

2 Related Work

Traditional NER methods are mainly implemented based on handcrafted rules, dictionaries, orthographic features or ontologies [32]. Further progress in this field has been achieved using statistical-based methods such as Hidden Markov Models and Conditional Random Fields (CRFs) [30]. In recent years, a large number of deep neural networks approaches have emerged, which greatly improved the recognition accuracy. Especially combinations of recurrent neural networks and CRFs have been successful [17]. The state-of-the-art methods show that word embeddings have a fundamental influence on the performance [21]. For a detailed overview of NER in the textual domain, see [32].

The extraction of information from document images has so far been rather a minor field of research in Document Image Analysis. However, the number of publications in recent years show that the interest in this topic has increased considerably [1,2,7,9,10,26,28]. Publications can be grouped according to their

[1] https://patrec.cs.tu-dortmund.de/cms/en/home/Resources/index.html

focus on machine-printed or handwritten document images. On modern machine-printed document images, recognition errors are usually so small that the application of NLP tasks can be considered solved. The situation is different with respect to historical documents, such as old newspapers. In these cases, recognition quality influences the results considerably and, therefore, there is still an active field of research [13,14]. Due to the high variability of handwriting, recognition of handwritten document images can generally be considered a more difficult problem compared to machine-printed ones. This problem is also true for the task of information extraction. Therefore, the current trend in contrast to approaches on machine-printed documents is to use the presumably more robust end-to-end architectures. The NER approaches on handwritten document images can be divided into segmentation free (i.e., entire document images are used without any segmentation) [7,10,9] and segmentation-based (i.e., a word or line level segmentation is assumed) [1,2,26,28]. One of the first approaches for NER on handwritten word images was provided by Adak et al. in [1]. In their approach, they use handcrafted features to decide whether a word image is a NE or not. Their approach only allows for detecting NEs and not for classifying them into predefined classes. At the International Conference on Document Analysis and Recognition in 2017, there was the Information Extraction in Historical Handwritten Records (IEHHR) competition [11]. It focuses on the automatic extraction of NEs on semi-structured historical handwritten documents. Approaches that initially performed recognition have won on both word and line segmentation. After the competition, Toledo et al. proposed two end-to-end architectures in [28] and evaluated them on the competition dataset. Their Bidirectional Long Short-Term Memory (BLSTM)-based approach is able to outperform the state-of-the-art results, presented in the competition. The end-to-end approach of Rowtula et al. [26] focuses on the extraction of NEs from handwritten documents containing unstructured text and has been both trained and evaluated on automatically generated NE tags for the IAM database. They observed that NEs are related to the position and distribution of Part-of-Speech (PoS) tags in a sentence. Therefore, they first train their model on PoS tag prediction using the CoNLL2000 dataset and synthetically generated word images. Then, they specialize the pre-trained model towards the real data, beginning with the prediction of PoS tags and lastly predicting the NE tags. Recently, Adak et al. propose in [2] an approach for word images from Bengali manuscripts. They extract patches from a single word image using a sliding window approach. Then, they extract a feature representation for each patch using their self designed convolutional architecture. The features are further encoded using a BLSTM. They apply attention weights in order to concentrate on the relevant patches. For each patch, a distribution over the NE classes is predicted and finally averaged across all patches. The highest score in the averaged distribution is then predicted as the NE class for the word image.

3 Datasets

The semi-structured dataset, used in the IEHHR competition (section 3.1), is so far the only established dataset in the field of NER on handwritten word images. Recently, Carbonell et al. evaluated two more datasets in [9], namely War Refugees and synthetic Groningen Meaning Bank (section 3.2). However, due to privacy reasons, only the synthetically generated dataset is publicly available. In order to enable evaluation on unstructured handwritten text, we created NE annotations for the George Washington dataset (section 3.3) and the IAM database (section 3.4). The main difference compared to the datasets in [1] and [26] is that the annotations were generated entirely manually, which avoids the errors caused by automatic taggers. In this section, we also show that the commonly known partitioning into training, validation and test set on GW and IAM are unsuitable for NER and present optimized splits. It is important to note that beside for the dataset in the IEHHR competition, all tag sets have a default class called O. This class is assigned to every word image that is not part of the predefined categories. Usually, there exist a huge class imbalance in every NE dataset with around 90 percent towards the O class.

3.1 Esposalles

The ESPOSALLES database [25] is an excerpt of a larger collection of historical handwritten marriage license books at the archives of the Cathedral of Barcelona. The corpus is written in old Catalan by a single writer in the 17th century. For the database, both line and word segmentations are available. The marriage records generally have a fixed structure, although there are variations in some cases. Therefore, the dataset can be considered semi-structured. For the IEHHR competition [11] 125 pages of this database were annotated with semantic information. There is an official partitioning into training and test data, containing 968 training and 253 test records. Each word is labeled with a category *(name, surname, occupation, location, state, other)* and a person *(husband, wife, husbands_father, husbands_mother, wifes_father, wifes_mother, other_person, none)*.

3.2 Synthetic Groningen Meaning Bank

The synthetic Groningen Meaning Bank (sGMB) dataset [9] consists of synthetically generated handwritten document pages obtained from the corpus of the Groningen Meaning Bank [8]. It contains unstructured English text mainly from a newspaper, whereby the words have been labeled with the following categories: *Geographical Entity, Organization, Person, Geopolitical Entity and Time indicator*. There is an official split containing 38048 training, 5150 validation and 18183 test word images. A possible disadvantage for the identification of NEs is the absence of punctuation marks in this dataset.

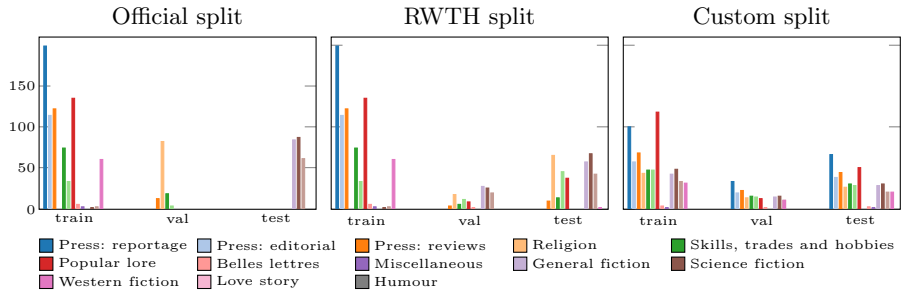

Fig. 2: The amount of document pages per genre in training, validation and test set. The histograms are shown for the official, RWTH and custom split of the IAM database.

3.3 George Washington

The George Washington (GW) dataset [23] has become the de-facto standard benchmark for word spotting. It consists of 20 pages of correspondences between George Washington and his associates dating from 1755. The documents were written by a single person in historical English. There is no publicly available semantic annotation for this dataset. Therefore, we created those manually and make them available for the community. The word images are labeled with the following categories: *Cardinal, Date, Location, Organization* and *Person*. Multiple splits have been published [5,24]. Unfortunately, they are unsuitable for our task, since NER is a sequence labeling problem and some sentences contain both training and validation words. Therefore, we present a more suitable split, which divides the documents into twelve training, two validation and six test pages. The partitioning was formulated as an optimization problem and solved using Answer Set Programming [18]. This involved dividing the pages from the dataset into training, validation and test data such that the NE categories are best split with respect to the ratio of 6:1:3. For this dataset, the transcription of a word image is only given in lowercase characters and does not contain punctuation. This could be a challenge for NER, since capitalization and punctuation are presumably important features for the identification of NEs.

3.4 IAM DB

The IAM Database [20] is a major benchmark for handwriting recognition and word spotting. The documents contain modern English sentences from the Lancaster - Oslo - Bergen Corpus [27] and were written by a total of 657 different people. The pages contain text from the genres listed in figure 2. The database consists of 1539 scanned text pages containing a total of 13353 text lines and 115320 words. The official partitioning splits the database in 6161 lines for training, 1840 for validation and 1861 for testing. These partitions are writer-independent, such

that each writer contributed to only one partition (either training, validation or test). As in the GW dataset, the official word-level partitioning unfortunately has the disadvantage that some sentences contain both training and validation data, which is unsuitable for NER. The official line segmentation does not have that problem. However, figure 3 shows that the distribution of text categories strongly differs from split to split. Therefore, the training data is not representative for the test and validation data. There is also another split specifically designed for handwriting recognition, referred to as RWTH split. Here the lines are partitioned into writer-independent training, validation and test partitions of 6161, 966 and 2915 lines, respectively. Even if the distribution of genres is considerably better compared to the official partitioning, figure 3 shows that it is a suboptimal split for NER.

Since it is essential for the training data to be representative for the test, which is not the case with the available splits, we propose a novel split. The partitioning was formulated as an optimization problem such that the documents from each text category are split as best as possible in the ratio of 3:1:2 between training, validation and test, while remaining writer-independent. Figure 3 shows that optimizing this criteria also improves the 3:1:2 ratio within the NE categories considerably. For annotating the IAM database, the same tag set was used as in OntoNotes Release 5.0 [22]. The tag set contains 18 categories that are well-defined in their published annotation guideline [2]. The categories are: *Cardinal, Date, Event, FAC, GPE, Language, Law, Location, Money, NORP, Ordinal, Organization, Person, Percent, Product, Quantity, Time* and *Work_of_art*. As there is a relatively small training set compared to the datasets used in the NLP domain, only a few examples exist for most categories. To overcome that problem, we also developed a smaller tag set that summarizes categories as best as possible and removes severely underrepresented categories. This resulted in a tag set consisting of only six categories: *Location (FAC, GPE, Location), Time (Date,Time), Cardinal (Cardinal, Ordinal, Percent, Quantity, Money), NORP, Person* and *Organization*. Furthermore, we use the official sentence segmentation of the dataset for all splits.

4 Two-staged Named Entity Recognition

For the comparison of end-to-end and two-stage approaches with respect to NER on segmented word images, we propose a two-stage approach. This approach is based on a state-of-the-art handwriting recognizer (HTR) and an NER model. The HTR was proposed by Kang et al. [15] and it is an attention-based sequence-to-sequence model for handwritten word recognition. It works on character-level and does not require any information about the language, except for an alphabet. The approach also has the advantage that it does not require any dataset-specific pre-processing steps. Therefore, the model produces satisfying results on most datasets and not only on a specific one. The NER model roughly follows the state-of-the-art architecture proposed by Lample et al. [17]. Here, the input words are

[2] https://bit.ly/3pyte8Q

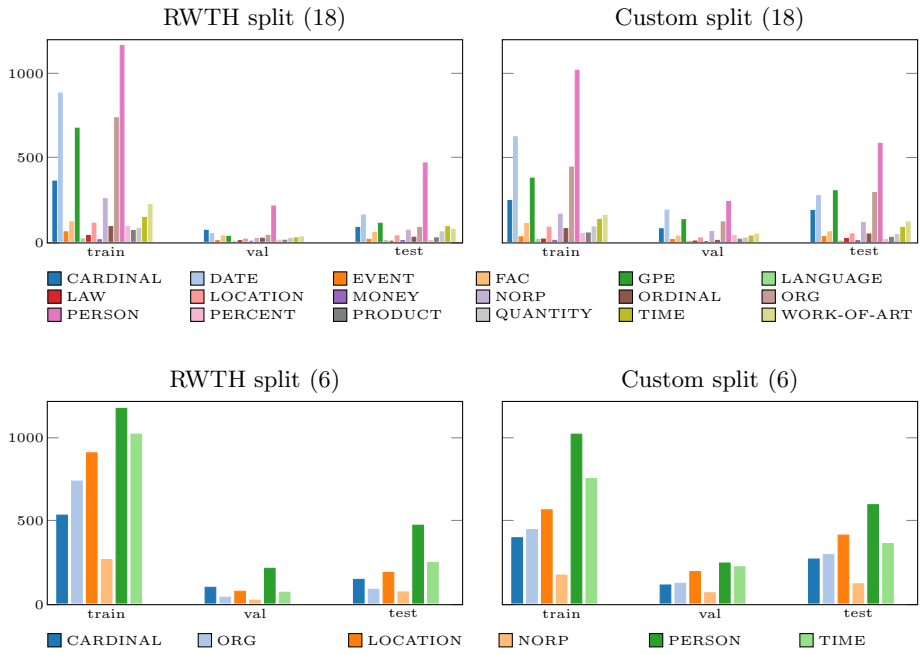

Fig. 3: The number of named entities in the training, validation and test partitions for the IAM database. Histograms are shown for the different combinations of splits (RWTH, Custom) and tag sets (six, eighteen).

first converted into a vector representation using a pre-trained word embedding model. For our datasets, the pre-trained RoBERTa base model of Huggingface [19] was shown to give the best results. Afterwards, these representations are encoded using a one layered BLSTM with a hidden layer size of 256. Finally, a CRF is used to predict NE tags for each word based on its encoding from the hidden layer. We implemented the NER model using the Flair framework [3].

The first step of our approach is to feed all word images into the recognizer in their order of occurrence on the document pages. If a sentence segmentation is available, the recognition results are divided accordingly. Otherwise, the entire page is defined as a single sentence. Finally, the sentences are processed sequentially by the NER model, which assigns a tag to each word image.

For the recognition model, we do not make any changes regarding the hyperparameters. We only customize the size of the input images, the maximum word length and the alphabet for each dataset. For the training of the NLP models in our approach, we use a mini batch size of 64. We update the model parameters using the SGD optimization procedure and the CRF loss. The learning rate is initially set to 0.1 and decreased by a factor of two whenever the F1-score on the validation data is not improving for five iterations. Since each word is to be

represented by exactly one vector and the RoBERTa model uses subword tokenization [19], we require a pooling strategy. This is especially important for the HTR results, because the RoBERTa model divides words that are not part of its vocabulary, known as Out-of-Vocabulary (OOV) words, into several subwords. Therefore, we represent a word as the average of all its subword representations. We also use a technique called scalar mix [21] that computes a parameterized scalar mixture of Transformer model layers. This technique is very useful, because for some downstream tasks like NER or PoS tagging it can be unclear, which layers of a Transformer-based model perform well. Another design decision is to use a trainable linear mapping on top of the embedding layer. This mapping ensures that the input to the BLSTM is learnable and does not come directly from the word embedding model. Given that the text in the Esposalles dataset is written in Spanish, we adapt the embedding approach by using a Spanish pre-trained flair embedding [4].

5 Experiments

In this section, we evaluate our two-stage approach on the four datasets introduced in section 3. We first describe the evaluation protocol in section 5.1. We then show the results in section 5.2 and compare our approach with two state-of-the-art end-to-end approaches for NER on segmented word images. For a fair comparison, we replicated the end-to-end models as best as possible and evaluate them with the same protocol and data. Finally, we discuss some potential methods for improving the robustness of our two-stage approach in section 5.3.

5.1 Evaluation Protocol

The F1-score is a suitable measure for evaluating NER models. However, there are several definitions of this measure, with macro and micro F1 being the most popular ones. In our experiments, we use the macro F1-score, which first computes the metric independently for each class and finally average these scores using the harmonic mean. Therefore, all classes are considered equally, preventing the score from being dominated by a majority class. It is important to note that we exclude the O class in our evaluation. The F1-score can be interpreted as a weighted average of the precision (P) and the recall (R) and is formally defined as shown in equation 1. Precision is the number of correctly predicted labels for a class divided by the number of all predicted labels for that class and recall is the number of correctly predicted labels for a class divided by the number of relevant labels for that class. Precision, recall and F1-scores are calculated per class and are finally averaged. It may happen that there is no element in the test set for a class, but the class was predicted for an element. In this case, the recall is to be defined as 0. It is also possible that there are no predicted labels for a class. In this case, the precision score should be set to 0.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{1}$$

Table 1: Handwriting recognition rates measured in Character Error Rate (CER) and Word Error Rate (WER) for the IAM, Esposalles, GW and sGMB datasets. For the IAM database, error rates are reported for both the RWTH and custom split.

Dataset	Dictionary free		Dictionary	
	CER	WER	CER	WER
IAM (RWTH)	6.80	17.67	6.20	10.70
IAM (Custom)	7.05	18.66	6.36	10.76
Esposalles	2.47	5.15	2.27	3.28
GW	5.24	14.52	4.10	6.05
sGMB	4.93	15.46	4.09	7.38

5.2 Results

In our comparison, we consider two state-of-the-art end-to-end models for NER on segmented word images. The first approach is the BLSTM-based model proposed by Toledo et al. [28]. We select this approach because it achieves state-of-the-art results on the IEHHR challenge dataset and it is able to outperform two-stage approaches. The other model was proposed by Rowtula et al. [26] and is currently the state-of-the-art approach on unstructured English word images.

In order to get an impression of the scores that can be achieved on the datasets and to obtain an indication of how the HTR errors affect the task, we also evaluate the NER model of our approach on perfect recognition results. For this purpose, the model receives the text annotations of the word images instead of the HTR results as input. In the following, we denote this approach as Annotation-NER.

As mentioned before, the HTR model used in our two-stage approach does not use any linguistic resources during recognition, such as language models or dictionaries. While this approach has the advantage of not penalizing OOV words, it also often helps with minor recognition errors. Since it is very unlikely to have no linguistic information about a given dataset in a real situation, we evaluate the other extreme case where there is a fixed vocabulary. For a dataset, this consists of its training, validation and test words. In the following, we denote our two-stage approach with HTR-NER if we do not use a dictionary and with HTR-D-NER otherwise.

The first step of our approach is to perform recognition for all word images in a given dataset. Table 1 shows the CERs and WERs of the test data for the four datasets. We report similar error rates as published in the literature [15]. Improvements are obviously possible with further optimizations and dataset-specific adaptations. However, this is not crucial for our comparison.

IAM The results in table 2 show that our custom split leads to a considerable performance increase in comparison to the RWTH split. This could be expected, as the training data is more representative regarding the validation and test sets. Our experiments show that the prediction of 18 categories constitutes a

Table 2: Named Entity Recognition performances for the IAM database measured in precision (P), recall (R) and macro-F1 (F1) scores. Results are shown for the different combinations of splits (RWTH, custom) and tag sets (six, eighteen).

Method	RWTH (6)			Custom (6)			RWTH (18)			Custom (18)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Annotation-NER	83.8	77.5	80.1	87.3	87.6	87.5	63.6	57.6	59.8	68.5	61.0	63.5
HTR-D-NER	78.6	73.0	75.4	83.7	78.7	81.0	60.4	50.9	54.2	62.5	53.3	56.3
HTR-NER	77.3	65.9	70.7	83.3	71.0	76.4	55.8	50.1	52.0	64.8	47.5	53.6
Rowtula et al. [26]	58.8	41.3	47.4	65.5	47.6	54.6	33.8	30.9	32.3	36.9	28.0	30.3
Toledo et al. [28]	45.3	28.8	34.0	50.2	31.4	37.4	26.4	10.8	14.9	35.4	13.4	18.0

hard task, resulting in low F1-scores. This is probably due to the small size of the training data, which leads to very few examples for some categories. Thus, the training set is not representative for the categories and the prediction of these is extremely difficult. Good results can be obtained with the tag set consisting of six categories and the Annotation-NER approach. Based on the results, it is also obvious that our two-stage approaches have large drops in comparison to the NER model working on the annotations. The additional use of a dictionary is able to reduce the errors in recognition and thus achieves better results for NER. The dictionary, thereby, greatly increases the recall scores, but shows similar performance in terms of precision. The two end-to-end approaches could only achieve comparatively low scores. We assume that the approach proposed by Toledo et al. performs rather poorly because it was developed for semi-structured data and cannot handle the strong imbalance on unstructured data. The model of Rowtula et al. is able to deliver considerably better scores on the IAM database. However, their method is optimized exactly for this dataset. The deviation between the scores in our analysis (see table 2) and their published results can be explained by their training and evaluation on different data and their consideration of the O class during evaluation. However, the most crucial point for the high scores in their publication is probably due to the use of automatic generated NE tags with spaCy. This is an NLP framework that uses the same features to predict NE and PoS tags and, therefore, presumably creates a much stronger correlation between the two tag types compared to manually labeled ones. Since Rowtula et al. exploit exactly this correlation in their approach and also pre-train on a large NLP PoS dataset, the differences between the scores are reasonable. In addition, they only use pages that meet a predefined sentence segmentation in both training and evaluation.

Esposalles The results for the dataset from the IEHHR challenge are shown in table 3a. As the IEHHR challenge does not only consist of the correct prediction of the two tags person and category, but also of the correct recognition of the text, we use our own evaluation protocol instead of the one used in the competition. The results show that the end-to-end as well as two-stage approaches perform

Table 3: Named Entity Recognition performances for the (a) Esposalles, (b) GW and (c) sGMB datasets measured in precision (P), recall (R) and macro-F1 (F1) scores. For Esposalles, the results are presented for each of the tag sets (person, category).

(a) Esposalles

Method	Person			Category		
	P	R	F1	P	R	F1
Annotation-NER	99.3	99.3	99.3	98.8	98.8	98.8
HTR-D-NER	99.3	99.2	99.3	98.5	98.2	98.3
HTR-NER	99.1	99.3	99.2	98.0	98.1	98.1
Rowtula et al. [26]	97.0	96.2	96.6	97.1	97.0	97.0
Toledo et al. [28]	98.5	97.8	98.1	98.5	97.8	98.1

(b) GW

Method	P	R	F1
Annotation-NER	96.5	84.7	89.6
HTR-D-NER	86.1	80.1	82.1
HTR-NER	86.9	78.3	81.3
Rowtula et al. [26]	76.4	59.8	66.6
Toledo et al. [28]	72.5	33.5	45.3

(c) sGMB

Method	P	R	F1
Annotation-NER	81.9	79.2	80.2
HTR-D-NER	81.8	75.9	78.4
HTR-NER	80.1	72.7	75.8
Rowtula et al. [26]	62.7	58.1	60.1
Toledo et al. [28]	44.3	35.3	38.8

well on semi-structured data. Thereby, all methods from our analysis are able to achieve comparable results.

GW Table 3b shows the results for the GW dataset. It is by far the smallest dataset and has few examples of each tag in the training set compared to the other datasets. However, the scope of context in the data is quite limited, making the training set highly representative for the validation and test data. This probably makes it possible to still obtain good results even under the limited amount of training material. Also, the end-to-end models achieve comparably good scores on this dataset.

sGMB Table 3c presents the results for the sGMB dataset and shows that the difference in F1-score between Annotation-NER and our two-stage approaches is small. A possible explanation could be that the NER model is optimized for segmented sentences and no sentence-level segmentation is available. In addition, the missing punctuation marks could also have a negative impact on the performance of the Annotation-NER model. Even though the difference in F1-scores between the Rowtula et al. approach and ours is smallest on this dataset, there is still an obvious difference.

5.3 Discussion

In this section, we discuss some potential methods for improving the robustness of a two-stage approach with respect to the task of NER on document images. In our experiments, we train the recognizer as well as the NLP model on the same training data, which leads to a sort of over-adaptation. As a result, the inputs to the NLP model have a considerably lower CER and WER during training compared to validation and test. This also implies that the NLP model has not been optimized for robustness with respect to HTR errors. Therefore, it is not surprising that the F1-score degrades considerably with increasing recognition errors for most datasets. We assume that the performance of our approach could improve when the training and test data have comparable errors with respect to recognition. Another possible improvement would be to adjust the pre-trained word embedding such that words and their erroneous HTR variants are close to each other in the vector space. Furthermore, HTR post-processing can presumably further close the gap between two-stage systems working on annotation and recognition results. However, it is quite remarkable that the NLP models without being specialized for HTR errors still perform better than models developed specifically for this task. We observe that NLP models work well under reasonable and easy to achieve recognition error rates, making two-stage approaches an interesting option. We do not state that the two-stage approach is generally more suitable for NER on word images compared to an end-to-end approach. However, our experiments show that there are currently more advantages for the two-stage models and they still show promising improvements in terms of tackling the task. In order to close the gap between image and text level, we believe that methods are needed that can provide semantic information of a word in an image. Furthermore, it must be possible to handle strongly unbalanced data during training. If it would be possible to adapt the advantages gained from the text level to the image domain, end-to-end approaches could be more appropriate again.

6 Conclusion

In this work, we propose and investigate a two-stage approach for Named Entity Recognition on word-segmented handwritten document images. In our experiments, we are able to outperform state-of-the-art end-to-end approaches on all four tested datasets, only using an unspecialized, standard handwriting recognizer from the literature and a textual Named Entity Recognition model. We demonstrate that due to the advantages from text level, two-stage architectures achieves considerably higher scores compared to end-to-end approaches for this task and have still potential for optimization. However, a final statement regarding the type of architecture for Named Entity Recognition on document images can not be derived based on our experiments. We also present and publish the first named entity tagged datasets on unstructured English text along with optimized splits as well as a suitable evaluation protocol.

References

1. Adak, C., Chaudhuri, B.B., Blumenstein, M.: Named entity recognition from unstructured handwritten document images. In: DAS. pp. 375–380. Santorini, Greece (2016)
2. Adak, C., Chaudhuri, B.B., Lin, C., Blumenstein, M.: Detecting named entities in unstructured Bengali manuscript images. In: ICDAR. pp. 196–201. Sydney, Australia (2019)
3. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL. pp. 54–59. Minneapolis, Minnesota (2019)
4. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING. pp. 1638–1649. Santa Fe, New Mexico, USA (2018)
5. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(12), 2552–2566 (2014)
6. Boros, E., Hamdi, A., Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., Doucet, A.: Alleviating digitization errors in named entity recognition for historical documents. In: CoNNL. pp. 431–441. Online (2020)
7. Boros, E., Romero, V., Maarand, M., Zenklová, K., Krecková, J., Vidal, E., Stutzmann, D., Kermorvant, C.: A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In: ICFHR. pp. 79–84. Dortmund, Germany (2020)
8. Bos, J., Basile, V., Evang, K., Venhuizen, N., Bjerva, J.: The groningen meaning bank. In: Proc. Joint Symposium on Semantic Processing. pp. 463–496 (2017)
9. Carbonell, M., Fornés, A., Villegas, M., Lladós, J.: A neural model for text localization, transcription and named entity recognition in full pages. Pattern Recognition, Letters **136**, 219–227 (2020)
10. Carbonell, M., Villegas, M., Fornés, A., Lladós, J.: Joint recognition of handwritten text and named entities with a neural end-to-end model. In: DAS. pp. 399–404. Vienna, Austria (2018)
11. Fornés, A., Romero, V., Baro, A., Toledo, J.I., Sánchez, J., Vidal, E., Lladós, J.: ICDAR2017 competition on information extraction in historical handwritten records. In: ICDAR. pp. 1389–1394. Kyoto, Japan (2017)
12. Gurjar, N., Sudholt, S., Fink, G.A.: Learning deep representations for word spotting under weak supervision. In: DAS. pp. 7–12 (2018)
13. Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., Doucet, A.: An analysis of the performance of named entity recognition over OCRed documents. In: Joint Conf. on Digital Libraries. pp. 333–334. Champaign, IL, USA (2019)
14. Hamdi, A., Jean-Caurant, A., Sidère, N., Coustaty, M., Doucet, A.: Assessing and minimizing the impact of OCR quality on named entity recognition. In: Int. Conf. on Theory and Practice of Digital Libraries. pp. 87–101. Lyon, France (2020)
15. Kang, L., Toledo, J.I., Riba, P., Villegas, M., Fornés, A., Rusiñol, M.: Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition. In: GCPR. pp. 459–472. Stuttgart, Germany (2018)
16. Krishnan, P., Jawahar, C.V.: Bringing semantics into word image representation. Pattern Recognition **108** (2020)
17. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: NAACL. pp. 260–270. San Diego, California, USA (2016)

18. Lifschitz, V.: What is answer set programming? In: Proc. AAAI Conf. on Artificial Intelligence. pp. 1594–1597. Chicago, Illinois, USA (2008)
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. ArXiv (2019)
20. Marti, U., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. IJDAR **5**(1), 39–46 (2002)
21. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAACL. pp. 2227–2237. New Orleans, Louisiana, USA (2018)
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: CoNNL. pp. 143–152. Sofia, Bulgaria (2013)
23. Rath, T.M., Manmatha, R.: Word spotting for historical documents. IJDAR **9**(2-4), 139–152 (2007)
24. Retsinas, G., Louloudis, G., Stamatopoulos, N., Gatos, B.: Efficient learning-free keyword spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1587–1600 (2019)
25. Romero, V., Fornés, A., Serrano, N., Sánchez, J., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition **46**, 1658–1669 (2013)
26. Rowtula, V., Krishnan, P., Jawahar, C.V.: PoS tagging and named entity recognition on handwritten documents. In: ICON. Patiala, India (2018)
27. Stig, J., Leech, G., Goodluck, H.: Manual of information to accompany the Lancaster-Oslo-Bergen Corpus of British English, for use with digital computers. http://korpus.uib.no/icame/manuals/LOB/INDEX.HTM (1978)
28. Toledo, J.I., Carbonell, M., Fornés, A., Lladós, J.: Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognition **86**, 27–36 (2019)
29. Tüselmann, O., Wolf, F., Fink, G.A.: Identifying and tackling key challenges in semantic word spotting. In: ICFHR. pp. 55–60. Dortmund, Germany (2020)
30. Wen, Y., Fan, C., Chen, G., Chen, X., Chen, M.: A survey on named entity recognition. In: Int. Conf. on Communications, Signal Processing, and Systems. pp. 1803–1810 (2019)
31. Wilde, M.D., Hengchen, S.: Semantic enrichment of a multilingual archive with linked open data. Digital Humanities Quarterly **11** (2017)
32. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: COLING. pp. 2145–2158. Santa Fe, New Mexico, USA (2018)