Exploring Semantic Word Representations for Recognition-free NLP on Handwritten Document Images

Oliver Tüselmann[0000−0002−8892−3306] and Gernot A. Fink[0000−0002−7446−7813]

Department of Computer Science, TU Dortmund University,
44227 Dortmund, Germany
{firstname.lastname}@cs.tu-dortmund.de

Abstract. A semantic analysis of documents offers a wide range of practical application scenarios. Thereby, the combination of handwriting recognizer and textual NLP models constitutes an intuitive solution. However, due to the difficulty of recognizing handwriting and the error propagation problem, optimized architectures are required. Recognition-free approaches proved to be robust, but often produce poorer results compared to recognition-based methods. In our opinion, a major reason for this is that recognition-free approaches do not use largely pre-trained semantic word embeddings, which proves to be one of the most powerful method in the textual domain. To overcome this limitation, we explore and evaluate several semantic embeddings for word image representation. We are able to show that context-based embedding methods are well suited for static word representations and that they are more predictive at word image level compared to classical static embedding methods. Furthermore, our recognition-free approach with pre-trained semantic information outperforms recognition-free as well as recognition-based approaches from the literature on several Named Entity Recognition benchmark datasets.

1 Introduction

Due to the combination of visual and textual properties, the semantic analysis of handwritten document images constitutes both an exciting and challenging field of research. Even though the focus of the Document Image Analysis community has been on visual rather than semantic tasks in the past, the community is steadily shifting towards the semantic analysis and understanding of document images [1,20,24,34,35,36]. Thereby, classical Natural Language Processing (NLP) tasks like Named Entity Recognition (NER) [1,38], Named Entity Linking [35] and Question Answering [24,36] have already been investigated for handwritten document images.

An intuitive approach for realizing NLP tasks on handwritten document images is to combine the advances from the visual and textual domain, using a two-stage model [38]. Thereby, a Handwritten Text Recognizer (HTR) transfers a given document into a textual representation and the outcome is processed

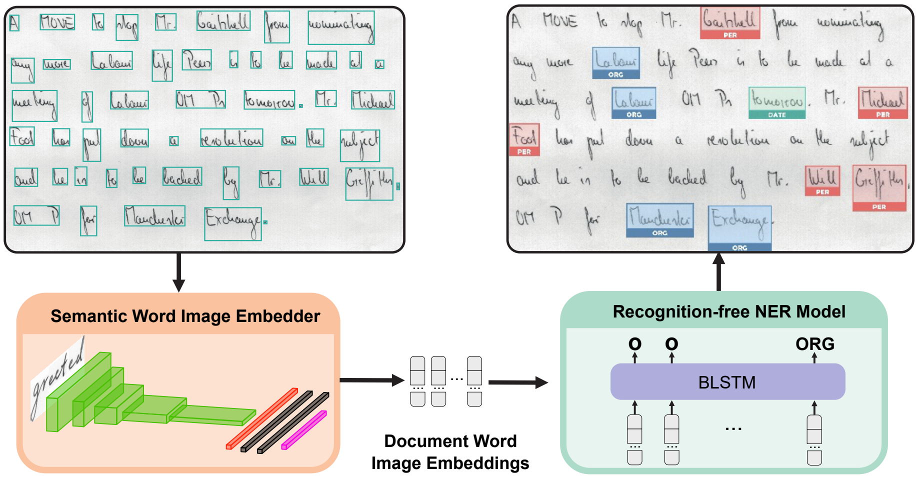
Semantic Word Image Embedder
greeted
Document Word Image Embeddings
Recognition-free NER Model
O
O
ORG
BLSTM
...
PER
ORG
DATE

Fig. 1: An Overview of our proposed recognition-free NLP approach on word-segmented handwritten document images with NER as the downstream task.

by a textual NLP model. Unfortunately, despite advances in machine learning, HTR approaches are still not perfect and can cause many recognition errors [38]. Several publications show that recognition errors have a strong negative impact on the performance of NLP models, mainly caused by error propagation [14,38]. To overcome this limitation, recognition-free end-to-end architectures are favored for documents that are difficult to recognize [24].

Even though recognition-free approaches can alleviate the error propagation problem, they are outperformed by two-stage recognition-based approaches on several semantic tasks [24,38]. In our opinion, this is mainly due to the fundamental drawback of not using pre-trained semantic word embeddings, which is one of the most powerful advantages of the NLP domain [40]. To overcome this limitation, we explore and evaluate which textually pre-trained semantic embeddings from the NLP domain are best suited for representing semantic information in word images. Furthermore, we incorporates these semantic embeddings into a recognition-free NLP framework for handwritten document images (see figure 1) and evaluate the performance on several NER datasets.

The remainder of this paper is organized as follows. Section 2 introduces related work in the fields of semantic word embeddings and NER on handwritten document images. In section 3, we present our recognition-free NLP framework and specifically focus on textually pre-trained semantic word embeddings for word image representation. We evaluate these representations and the framework for NER on handwritten document images in section 4. Finally, we summarize our results in section 5.

2 Related Work

This section reviews related work regarding the main concepts used in our proposed recognition-free NLP framework. We provide an overview of syntactic and semantic word embedding methods and show how they are predicted from word image level. We further present related work in the field of NER on document images.

2.1 Word Embeddings

Processing textual words using electronic devices, requires a transformation of these words into numeric representations. Current methods realize such a transformation by using word embeddings. They find their application throughout all NLP tasks and many other domains [31]. Thereby, the use of specialized embedding techniques lead to a significant performance improvement in a wide variety of areas, including NLP [31] and Document Image Analysis tasks [24,32,36]. Even though there are numerous embedding methods, we will only consider semantic and syntactic word embedding approaches in the following.

The majority of semantic word embedding approaches are based on the distributional hypothesis [15]. This hypothesis states that words occurring in similar contexts tend to have similar meanings. Approaches can be roughly divided into static [4,25] and context-based methods [2,8,27]. Static approaches generate embeddings independently of their context and thus map a word always to the same vector representation [4,25]. These methods have the fundamental drawback of ignoring the fact that a word can have various meanings in different contexts. In recent years, several context-based embeddings approaches have been published [2,8,27]. These approaches are trained on language modeling tasks and rely on recurrent neural networks [2,27] or transformer-based architectures [8]. The change from static to context sensitive embeddings led to better results in almost all tasks in the NLP domain [10]. For a detailed overview of semantic word embeddings in the textual domain, see [31].

While semantic information refers to the meaning of a word, syntactic information represents its structural properties. Even though syntactic word embeddings seem to have a minor importance in the field of textual semantic analysis tasks, they are commonly used in the Document Image Analysis domain [32,34,36]. Syntactic word embeddings (e.g. Pyramidal Histogram of Character [3]) are often used in the field of handwritten word images to allow a similarity comparison between a textual query and a word image [3,32,34].

2.2 Word Image Mapping

Currently, methods based on Convolutional Neural Networks (CNNs) are most suitable for obtaining semantic and syntactic word embeddings at the word image level [20,37,39]. A variety of approaches have been published for realizing a syntactic representation on word image level [19,32,39]. Whereas semantic embedding approaches follow a unified strategy by predicting textually pre-trained embeddings for word images [20,34,37,39]. First approaches in this area map word images into a textually pre-trained semantic space by using a two-stage CNN-based approach [37,39]. Thereby, the word images are converted into a feature representation and afterwards mapped into the semantic space. End-to-end approaches are able to outperform two-stage architectures on semantic word image mapping [20,34]. Recently, the realization of a combined syntactic and semantic word image representation has been investigated [20,34].

2.3 Named Entity Recognition

Named Entity Recognition (NER) is a sequence labeling task with a long tradition in NLP [40]. The goal of this task is to extract named entities (e.g. places, person, organizations) from an unstructured text. Traditional approaches mainly rely on handcrafted rules, dictionaries or ontologies [40]. Today, methods using neural architectures outperform traditional ones [2,8,21]. Especially, the combination of a Long Short-Term Memory (LSTM) and a Conditional Random Field (CRF) yields state-of-the-art scores on many benchmarking datasets [21]. Similar to many other NLP tasks, the use of pre-trained semantic word embeddings leads to a considerable performance gain on most benchmarks [2,8]. For a detailed overview of NER in the textual domain, see [40].

There is a wide range of applications in the field of NER on document images. In the following we focus on approaches that work directly on word image level and not on already transcribed text. Publications in this field can be grouped according to their focus on machine-printed [9,14] and handwritten document images [1,30,33,38]. A further categorization of the works can be made on the basis of segmentation-free [6,11] and segmentation-based [1,30,33] approaches. Thereby, segmentation-free approaches work on the entire document image, whereas segmentation-based approaches require a line or word segmentation. A combination of a CNN and an LSTM has proven to be particularly successful for segmentation-based NER approaches [1,30,33]. Furthermore, it has been shown that integrating additional information (e.g. part-of-speech tags) [30] or using an attention mechanism [1] can lead to further improvements in this domain. Tueselmann et al. showed recently, that a two-stage architecture consisting of an HTR and a textual NER model is able to outperform end-to-end approaches on several NER datasets [38].

3 Method

In this section, we present our recognition-free NLP framework for word segmented handwritten document images (see figure 1). The approach consists of a textually pre-trained semantic word embedding, a word image mapper and a recognition-free NLP model. Thereby, the word image mapper processes the word images in the order in which they occur on a pre-segmented document image and predicts a semantic word embedding for each of them. Afterwards, these embeddings are transferred to a recognition-free NLP model (e.g. NER), which fulfills the appropriate task. This framework closely follows the two-stage recognition-based approach as proposed in [38], however, we avoid an explicit recognition step and obtain the semantic word representations directly on word image level.

3.1 Semantic Word Embeddings

Semantic word embeddings play an important role in tasks related to text understanding and lead to considerable improvements in almost all areas of NLP [31]. Especially, context-based approaches achieved major performance gains [8,27]. In the field of handwritten document image analysis, however, only static word embeddings have been used so far [20,37,39]. The main reason for this is most probably that already the mapping of context-independent embeddings poses a major challenge [37]. Recently, Ethayarajh showed in [10] that contextualized semantic representations (e.g. BERT) contain powerful types of context-independent embeddings in their first layers. These representations are able to outperform traditional context-independent approaches on many static semantic benchmarks [10]. Given these new insights, we evaluate in this work whether these outcomes can be transferred to the word image domain. Furthermore, we investigate which word embedding approaches from the textual domain are best suited for obtaining a powerful semantic word image representation. In the following, we provide a short overview of word embedding methods that we consider in our evaluation.

For our recognition-free NLP framework, we evaluate static [4,17,26] as well as contextualized [2,8,27] semantic embedding approaches. A classical static method is GloVe [26] which determines its semantic representations by using coincidence statistics between a target word and its context words defined by a fixed context window. This approach has the major disadvantage of being unable to predict embeddings for words that were not part of the training. To overcome this limitation, subword-based approaches like FastText (FT) [4] and BytePair [17] have been published which split words into subwords and combine their embeddings into a single representation. The drawback of static methods is that the word order is not taken into account. Context-based methods are used to encode this type of information. The training of these models focuses on language modeling. First approaches like ELMO [27] and Flair [2] use LSTM-based architectures. A fundamental difference between these two approaches is that Flair processes the textual input purely character based while ELMO uses a mixture of character and static word embeddings. State-of-the-art methods like BERT [8] are based on transformers and subword-based representations. Furthermore, we consider combinations of semantic representations in our evaluation, as they often lead to performance improvements in the textual domain [13].

3.2 Word Image Representation

For obtaining semantic word image representations, we use the same modified ResNet architecture (Attribute-ResNet) as proposed in [34]. The Attribute-

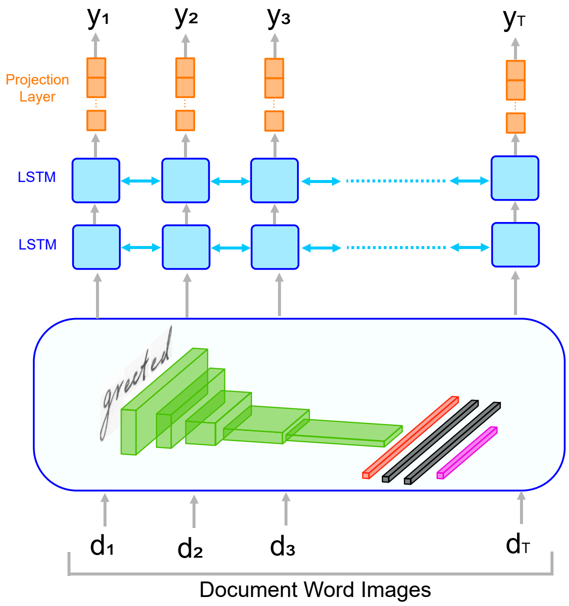

Fig. 2: Our proposed architecture for realizing a robust recognition-free NER system incorporating semantic information.

ResNet uses a ResNet34 architecture [16] for feature extraction, whereby the global average pooling layer at the end of the network is replaced with a Temporal Pyramid Pooling (TPP) layer. The output of the TPP layer is transferred into a three-layered Fully-Connected Network (FCN). This FCN has as many neurons in the last layer as there are dimensions in the word representation to be predicted (e.g. FastText = 300). Except for the final layer, the ReLU activation function is applied to the output of all layers in the network.

3.3 Named Entity Recognition

The NER approach roughly follows the architecture proposed by Toledo et al. [33]. Figure 2 provides an overview of our model. The first step of our approach is the prediction of semantic word image representations for each word image from the document $(d_1, ..., d_T)$. We further capture relations among these representations by using a two-layered Bidirectional-LSTM (BLSTM). Finally, a linear layer is applied to each hidden layer of the BLSTM in order to obtain a named entity tag for each word image $(y_1, ..., y_T)$.

4 Experiments

We evaluate the semantic quality of word embeddings for handwritten word images by using an efficient strategy from the textual NLP domain, which consists of an intrinsic and an extrinsic evaluation [29]. In this context, an intrinsic evaluation involves tasks that are simple and fast to compute and allows inference about the performance on real-world tasks. An extrinsic evaluation, on the other hand, focuses on the actual task (e.g. NER, QA) and is thus more time-consuming.

For our intrinsic and extrinsic experiments, we describe the evaluation datasets, implementation details as well as evaluation protocols. We further present and discuss the results of the two evaluations in this section.

4.1 Datasets

For our experiments, both intrinsic and extrinsic evaluation datasets are required. In order to compare with approaches from the literature, we use the IAM-DB, GNHK and sGMB datasets for our intrinsic evaluation. Similar to [38], we use the IAM-DB, sGMB, and George Washington datasets for our extrinsic evaluation. Moreover, the HW-Synth dataset is used for pre-training the word image mapper.

IAM-DB The IAM Database [23] is a major benchmark for a variety of handwritten document image tasks. The documents contain modern English sentences written by a total of 657 different people. The database consists of 1539 scanned text pages containing a total of 13353 text lines and 115320 words. Tueselmann et al. manually annotated the dataset with named entity labels and proposed an optimized semantic split into train, validation and test data [38]. There are two versions of this dataset available with different label sets containing 6 and 18 classes.

HW-Synth The HW-Synth (HW) dataset [18] provides a collection of synthetically rendered word images. The dataset is often used for pre-training handwritten models. The word images are generated by True Type Fonts that resemble handwriting. The vocabulary consists of the 12000 most common words from the English language. For each word, 50 training and 4 test images are generated. The font is randomly sampled from over 300 publicly available fonts.

GNHK The GoodNotes Handwriting Kollection (GNHK) dataset [22] includes unconstrained camera-captured document images of English handwritten notes. It consists of 687 documents containing a total of 9363 text lines and 39026 words. The official partitioning divides the data into training and test sets with a ratio of 75% and 25%, respectively.

SGMB The synthetic Groningen Meaning Bank (sGMB) dataset [6] consists of synthetically generated handwritten document pages obtained from the corpus of the Groningen Meaning Bank [5]. The dataset provides unstructured English text and splits the data into 38048 training, 5150 validation and 18183 test word images. The label set consists of the following categories: *Geographical Entity, Organization, Person, Geopolitical Entity and Time indicator*.

George Washington The George Washington (GW) dataset [28] 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. The word images are labeled with the following categories: *Cardinal, Date, Location, Organization* and *Person*.

4.2 Implementation Details

The semantic network follows the same training and optimization strategy as described in [34]. To obtain gold standard semantic embeddings for our word images, we used the Flair framework [2]. Thereby, we used the uncased, base model of BERT and the default English models for ELMO, BytePair and GloVe. For the Flair embeddings, the pre-trained forward and backward English models are used and for FastText the Common Crawl English model [12]. Furthermore, the PHOC representation consists of layers $2, 3, 4, 5$ and an alphabet with characters $a - z$ and $0 - 9$. It is important to note that for all embeddings, we have lowercased the transcriptions and followed the same alphabet as used for PHOC. In our experiments we realize a combined representation of semantic approaches by concatenating their embeddings.

The BLSTM model of our NER architecture uses a hidden layer size of 256 and a dropout of 0.5. For optimization we use the Cross Entropy Loss and the ADAM optimizer. The learning rate is initially set to 0.001 and divided by two whenever the training loss does not decrease in a certain range within 10 epochs. We follow the label smoothing approach proposed by [7]. There is no sentence segmentation and all word images of a document are processed simultaneously.

4.3 Evaluation Protocol

Since we evaluate the use of various textual semantic embeddings for word image representation intrinsically as well as extrinsically, several metrics and protocols are required. For this purpose, we use syntactic and semantic metrics for our intrinsic evaluation and NER task for our extrinsic evaluation.

Intrinsic Evaluation For an intrinsic evaluation of the word image representation methods introduced in section 3.1, a semantic as well as syntactic metric is required. We use the exact same metrics and protocols as described in [20,34]. Thereby, word spotting [3,20,32] is used as the syntactic and Word Analogy (WA) [25] as the semantic quality measure. Word spotting is a retrieval-based

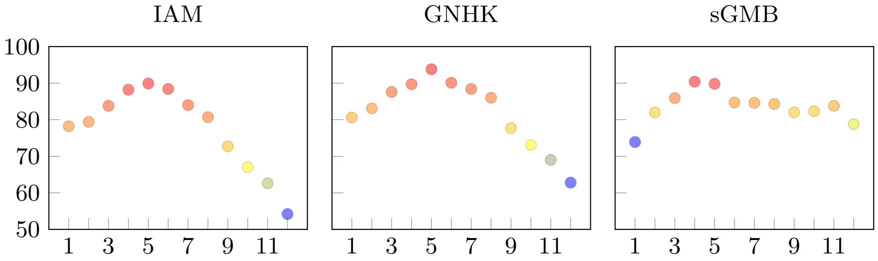

Fig. 3: Inspecting the quality of each layer (1-12) in the BERT model for use as static word embedding on the IAM, GNHK and sGMB dataset. The quality is determined using the WA score measured in accuracy [%].

task, which obtains a ranking of word images from a collection of document images based on its similarity w.r.t. a given query. There exists a variety of different query types with Query-by-Example (QbE) and Query-by-String (QbS) being the most prominent ones. In QbE applications, the query is a word image, whereas in QbS it is a textual string representation. Mean Average Precision (mAP) is the de-facto standard metric for evaluating retrieval tasks.

In the WA task, three words a, b and c are given and the goal is to infer the fourth word d that satisfies the following condition: a is to b as c is to d. We use the collection of human-defined WA examples proposed in [25]. Note, that questions which contain words that are not part of the test corpus of a dataset are excluded from the evaluation. The accuracy of correctly predicted analogies is used as the final semantic evaluation score.

Named Entity Recognition We use the macro F1-score with the exact same protocol as described in [38]. The F1-score can be interpreted as a weighted average of precision (P) and recall (R) and is formally defined as shown in equation 4.3. In macro F1 the precision, recall and F1-scores are calculated per class and are finally averaged. It is important to note that we exclude the non-entity (O) class in our evaluation.

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

4.4 Intrinsic Evaluation

We evaluate the capability of various textual semantic word representations to represent semantic in word images. For this purpose, we first present and evaluate our method for extracting static embeddings from context-based approaches. Afterwards, we determine the quality of the semantic word representations introduced in section 3.1 on the gold standard annotations of each dataset using WA. Finally, we evaluate the prediction of semantic word representations at word-image level both semantically and syntactically.

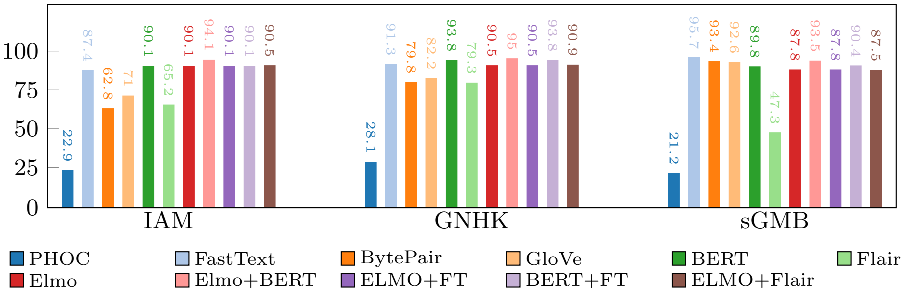

Fig. 4: WA scores for different word embedding methods from the NLP domain on the gold standard annotations of the IAM, GNHK and sGMB datasets. The results are given in accuracy [%].

Table 1: Performances on the four evaluated datasets using accuracy [%] for the WA task (semantic) and mAP for QbE and QbS word spotting (syntactic).

Method	IAM			GW			sGMB			GNHK		
	QbE	QbS	WA	QbE	QbS	WA	QbE	QbS	WA	QbE	QbS	WA
PHOC	91.9	96.2	23.9	96.7	96.8	-	95.7	94.2	20.3	81.5	81.8	28.1
FastText (FT)	86.5	72.0	80.5	95.3	79.5	-	89.7	63.5	75.6	75.2	53.2	70.2
BytePair	87.0	72.2	58.6	94.8	82.2	-	94.2	71.0	85.7	73.4	50.7	60.3
GloVe	87.1	72.2	67.7	96.2	78.6	-	95.0	71.3	85.9	76.9	53.7	66.1
BERT	89.2	74.8	85.1	96.6	81.3	-	95.4	74.6	79.6	77.6	55.3	67.4
Flair	87.4	85.8	49.2	94.7	92.4	-	95.8	82.5	35.8	77.2	67.2	38.4
ELMO	87.5	78.5	86.8	96.4	91.1	-	94.5	75.9	78.6	74.7	58.6	73.6
ELMO + BERT	88.5	78.9	**88.9**	92.7	81.5	-	94.2	76.0	77.9	77.7	61.1	**77.3**
ELMO + FT	87.3	78.6	87.4	96.2	90.5	-	94.4	75.7	77.7	78.1	62.0	76.9
BERT + FT	88.4	74.3	85.1	95.7	83.6	-	95.5	74.2	79.1	78.8	57.8	60.3
ELMO + Flair	90.2	85.0	74.8	96.3	93.7	-	94.6	77.9	54.1	78.0	65.8	55.0

For obtaining static embeddings from context-based approaches, we utilize the findings of Ethayarajh [10] and use the layer from the context-based model that provides the best static characteristics. Figure 3 visualizes the word analogy scores of each layer within the context-based BERT model. The results show that the performances of the individual layers differ considerably. Thereby, the first layers seem to be able to realize a powerful static word representation. Whereby, the fifth layer proves to be most suitable due to its performance on all three datasets. From the fifth layer onwards, the quality decreases and the last layers seem to be rather context-sensitive and thus poorly represent static information.

Figure 4 visualizes the WA scores for our considered semantic word representations on the gold standard annotations of the three datasets introduced in section 4.1. The static embeddings extracted from the context-based approaches (BERT, ELMO) clearly demonstrate improved or similar scores compared to the static approaches (FastText, GloVe, BytePair). The combination of semantic embeddings seems to be promising, especially the combination of the ELMO and BERT embedding achieves good results on all datasets. Flair is a purely

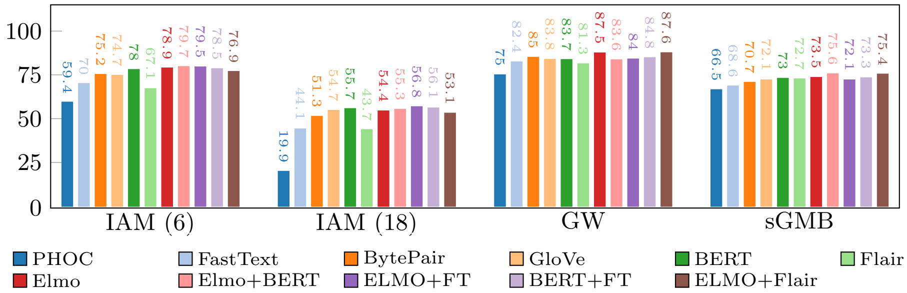

Fig. 5: NER results for predicted word embeddings at the word image level. We report the macro F1 scores [%] for the examined semantic word embeddings on the four evaluation datasets.

character-based embedding and leads to comparatively low scores in our evaluation.

Interestingly, the static approaches lead to high WA scores on the sGMB dataset. That is primarily due to the different examples in the WA task for each dataset, since only examples are considered in which the result of the analogy occurs as a word image in the test set. Since the sGMB dataset consists of several news texts, the analogies comprise more than 90% of pure relations between countries and cities. Those relations seem to be very well encoded in the static embeddings (FastText, BytePair, GloVe). This raises the question regarding the usefulness of intrinsic metrics for evaluating semantic quality and whether the focus should rather be on downstream tasks when evaluating semantic representations.

The results obtained for predicting the semantic embeddings at word image level generally follow the trends observed in the WA scores on the textual gold standard data. The BERT and ELMO representations improved the QbS and QbE scores and thus encode better syntactic information. Especially, the ELMO embedding appears to be much more suitable based on its mixture of character and word representation. Flair can achieve high syntactic scores, however, the performance on the semantic evaluation measure is quite low. Similar to the gold standard annotation, the combination of ELMO and BERT is able to achieve high semantic scores on almost all datasets.

4.5 Extrinsic Evaluation

We use the challenging and well-known NER task for our extrinsic evaluation. Figure 5 provides the performances of our intrinsically evaluated semantic embeddings on several NER datasets measured in macro F1 score [%]. The embeddings used so far in the literature for building recognition-free NLP approaches (PHOC and FastText) perform rather poorly on these datasets compared to our newly introduced semantic representations. While the Flair embedding can only achieve comparatively low values particularly on the IAM dataset, the BERT

Table 2: NER performances for the evaluated datasets measured in precision (P), recall (R) and macro-F1 (F1) scores.

Method	IAM (6)			IAM (18)			GW			sGMB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Annotation-NER [38]	87.3	87.6	87.5	68.5	61.0	63.5	96.5	84.7	89.6	81.9	79.2	80.2
HTR-NER [38]	83.3	71.0	76.4	64.8	47.5	53.6	86.9	78.3	81.3	80.1	**72.7**	**75.8**
Rowtula et al. [30]	65.5	47.6	54.6	36.9	28.0	30.3	76.4	59.8	66.6	62.7	58.1	60.1
Toledo et al. [33]	50.2	31.4	37.4	35.4	13.4	18.0	72.5	33.5	45.3	44.3	35.3	38.8
Ours (ELMO+BERT)	**86.4**	**74.6**	**79.7**	**78.1**	**51.2**	**55.3**	**96.2**	79.5	**83.0**	**80.6**	72.0	75.6

and ELMO representations achieve good performances. Especially the combination of semantic embeddings proves to be promising and leads to the highest scores on all datasets. There is a correlation between the WA scores from the intrinsic evaluation and the F1 scores achieved on the NER task, however, it is not possible to generally conclude that a higher WA score leads to improved results on the downstream task.

To compare our recognition-free NER model with approaches from the literature, we use a combination of ELMO and BERT as the semantic representation. The results are shown in table 2. Thereby, Annotation-NER is a recognition-based approach that works on the gold standard annotations of the datasets and thus reflects the NER performances under perfect recognition. The results show that our approach obtains considerably superior scores compared to the recognition-free approaches from the literature ([30,33]). This demonstrates the importance of using pre-trained semantic information. Moreover, except on the sGMB dataset, our approach is able to outperform the purely recognition-based approach of [38] (HTR-NER). Thereby, our approach obtains a similar performance on the sGMB dataset and the recognition-based approach benefits from low recognition errors due to the synthetic nature of this dataset. In the case of the IAM dataset, it should be noted that the word image mapper was pre-trained on the word spotting split of the dataset and thus a potential test set leak could exist.

4.6 Discussion

Further interesting research questions are, what is the best way to incorporate semantic information into our architecture and whether this information is beneficial. For this purpose, we examine three approaches. The first approach is the same as in the previous sections. Here, we train the word mapping network separately from the downstream task and subsequently freeze the pre-trained network while training on the downstream tasks. Thus, the parameters in the Attribute-ResNet are not adjusted during the training process. The second approach also trains the semantic model separately, however, during training of the downstream task, the parameters of the Attribute-ResNet can be adjusted. The last approach is an end-to-end approach, which does not rely on a seman-

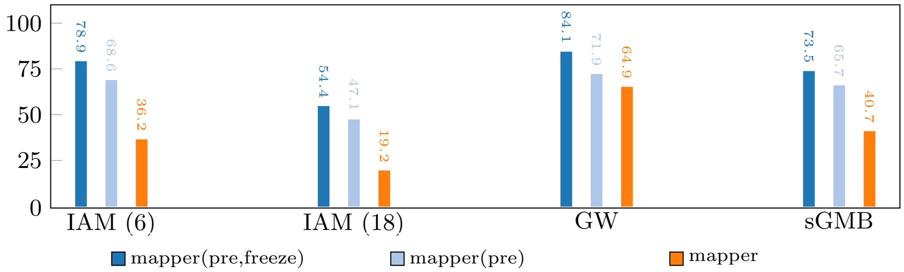

Fig. 6: Examine whether the pre-training (pre) of the image mapper (mapper) is helpful and how to integrate it most effectively into the NLP model.

tically pre-trained network and is similar to the approach of [33]. Whereas the Attribute-ResNet is used instead of the PHOCNet [32].

The results clearly show that a pre-training of the Attribute-ResNet is extremely important. Furthermore, the results show that changing the parameters of the Attribute-ResNet during the training of the downstream task is counterproductive. This is probably due to the fact that the datasets are quite small and thus quickly lead to overfitting when the large number of parameters in the ResNet are adjustable.

5 Conclusions

In this work, we present a recognition-free framework for NLP tasks on word-segmented handwritten document images. Our approach focuses on the prediction of textually pre-trained semantic embeddings for word images. For this purpose, we intrinsically evaluated both static and context-based approaches and demonstrate that the context-based approaches and especially their combination are often more suitable than the previously used static embeddings such as FastText. In our extrinsic evaluation on several Named Entity Recognition datasets, we can support the findings from the intrinsic evaluations and show that our approach can outperform both recognition-free as well as recognition-based approaches from the literature.

References

1. Adak, C., Chaudhuri, B.B., Lin, C., Blumenstein, M.: Detecting named entities in unstructured Bengali manuscript images. In: Proc. Int. Conf. on Document Analysis and Recognition. pp. 196–201. Sydney, Australia (2019)
2. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proc. Int. Conf. on Computational Linguistics. pp. 1638–1649. Santa Fe, NM, USA (2018)
3. 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)

4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
5. Bos, J., Basile, V., Evang, K., Venhuizen, N., Bjerva, J.: The Groningen meaning bank. In: Proc. Joint Symposium on Semantic Processing. pp. 463–496. Trento, Italy (2013)
6. 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)
7. Carbonell, M., Villegas, M., Fornés, A., Lladós, J.: Joint recognition of handwritten text and named entities with a neural end-to-end model. In: Int. Workshop on Document Analysis Systems. pp. 399–404. Vienna, Austria (2018)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Annual Conf. of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186. Minneapolis, MN, USA (2019)
9. Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers. In: European Conf. on Information Retrieval. pp. 524–532. Lisbon, Portugal (2020)
10. Ethayarajh, K.: How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Proc. Conf. on Empirical Methods in Natural Language Processing. pp. 55–65. Hong Kong (2019)
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: Proc. Int. Conf. on Document Analysis and Recognition. pp. 1389–1394. Kyoto, Japan (2017)
12. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proc. Int. Conf. on Language Resources and Evaluation. Miyazaki, Japan (2018)
13. Gupta, P., Jaggi, M.: Obtaining better static word embeddings using contextual embedding models. In: Joint Conf. of the Annual Meeting of the Association for Computational Linguistics and the Int. Joint Conf. on Natural Language Processing. pp. 5241–5253. Bangkok, Thailand (2021)
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. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conf. on Computer Vision and Pattern Recognition. pp. 770–778. Las Vegas, NV, USA (2016)
17. Heinzerling, B., Strube, M.: BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In: Proc. Int. Conf. on Language Resources and Evaluation. Miyazaki, Japan (2018)
18. Krishnan, P., Jawahar, C.V.: Generating synthetic data for text recognition. CoRR **abs/1608.04224** (2016)
19. Krishnan, P., Jawahar, C.V.: HWNet v2: An efficient word image representation for handwritten documents. Int. Journal on Document Analysis and Recognition **22**, 387–405 (2019)
20. Krishnan, P., Jawahar, C.V.: Bringing semantics into word image representation. Pattern Recognition **108**, 107542 (2020)

21. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Annual Conf. of the North American Chapter of the Association for Computational Linguistics. pp. 260–270. San Diego, CA, USA (2016)
22. Lee, A.W.C., Chung, J., Lee, M.: GNHK: A dataset for English handwriting in the wild. In: Proc. Int. Conf. on Document Analysis and Recognition. pp. 399–412. Lausanne, Switzerland (2021)
23. Marti, U., Bunke, H.: The IAM-database: An English sentence database for offline handwriting recognition. Int. Journal on Document Analysis and Recognition **5**(1), 39–46 (2002)
24. Mathew, M., Gómez, L., Karatzas, D., Jawahar, C.V.: Asking questions on handwritten document collections. Int. Journal on Document Analysis and Recognition **24**, 235–249 (2021)
25. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Int. Conf. on Learning Representations. Scottsdale, AZ, USA (2013)
26. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proc. Conf. on Empirical Methods in Natural Language Processing. pp. 1532–1543. Doha, Qatar (2014)
27. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Annual Conf. of the North American Chapter of the Association for Computational Linguistics. pp. 2227–2237. New Orleans, LA, USA (2018)
28. Rath, T.M., Manmatha, R.: Word spotting for historical documents. Int. Journal on Document Analysis and Recognition **9**(2-4), 139–152 (2007)
29. Resnik, P., Lin, J.: The Handbook of Computational Linguistics and Natural Language Processing, chap. 11, pp. 271–295 (2010)
30. Rowtula, V., Krishnan, P., Jawahar, C.V.: PoS tagging and named entity recognition on handwritten documents. In: Int. Conf. on Natural Language Processing. Patiala, India (2018)
31. Sezerer, E., Tekir, S.: A survey on neural word embeddings. CoRR **abs/2110.01804** (2021)
32. Sudholt, S., Fink, G.A.: PHOCNet: A deep convolutional neural network for word spotting in handwritten documents. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition. pp. 277—282. Shenzhen, China (2016)
33. 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)
34. Tüselmann, O., Brandenbusch, K., Chen, M., Fink, G.A.: A weighted combination of semantic and syntatic word image representations. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition. pp. 285–299. Hyderabad, India (2022)
35. Tüselmann, O., Fink, G.A.: Named entity linking on handwritten document images. In: Int. Workshop on Document Analysis Systems. pp. 199–213. La Rochelle, France (2022)
36. Tüselmann, O., Müller, F., Wolf, F., Fink, G.A.: Recognition-free question answering on handwritten document collections. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition. pp. 259–273. Hyderabad, India (2022)
37. Tüselmann, O., Wolf, F., Fink, G.A.: Identifying and tackling key challenges in semantic word spotting. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition. pp. 55–60. Dortmund, Germany (2020)

38. Tüselmann, O., Wolf, F., Fink, G.A.: Are end-to-end systems really necessary for NER on handwritten document images? In: Proc. Int. Conf. on Document Analysis and Recognition. pp. 808–822. Lausanne, Switzerland (2021)
39. Wilkinson, T., Brun, A.: Semantic and verbatim word spotting using deep neural networks. In: Proc. Int. Conf. on Frontiers in Handwriting Recognition. pp. 307–312. Shenzhen, China (2016)
40. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: Proc. Int. Conf. on Computational Linguistics. pp. 2145–2158. Santa Fe, NM, USA (2018)