Weakly Supervised Information Extraction from Semi-Structured Document Images

Fabian Wolf[1,2], Oliver Tüselmann[1,2], Christoph Rass[3], Gernot A. Fink[1,2]
[1]Department of Computer Science, TU Dortmund University, Germany
Email: {firstname.lastname}@cs.tu-dortmund.de
[2]LAMARR, Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany
[3]Chair of Modern History and Historical Migration Studies, Osnabrück University, Germany
Email: christoph.rass@uni-osnabrueck.de

I. Introduction

Throughout the 20th century, the modern state produced semi-structured files and forms to organize and handle rapidly expanding amounts of data with increasing speed and efficiency. In Germany, massive card file indices, often created to store person-related data, spread and grew into massive and ubiquitous datafication systems. As historical sources, the surviving archival holdings of semi-structured mass data have long been used for qualitative research in small samples or case studies. Making large collections of Thousands or even Millions of forms or index cards machine readable was not feasible by manual transcription. Recent developments in automated recognition technologies that can recognize machine-written and handwritten text in semi-structured historical files promise to change that and open up entire collections to systematic data extraction and in-depth indexing. While traditional approaches rely on the manual creation of labeled training sets, we investigate an *annotation-free* process that does not rely on manually annotated data.

II. Dataset

Pilot projects by the University of Osnanbrück's research group on modern history and historical migration studies in cooperation with the State Archive of Lower Saxony and external partners have demonstrated that reliable data extraction from card file indices can be achieved. Based on initial results, our research consortium has set out to make refined recognition tools available to researchers and archives. To this end, we revisit the Gestapo card file index preserved at the Osnabrück branch of the State Archive, one of only a handful that has survived. This index was created by the Prussian political police in the late 1920s as part of the information technology revolution of the interwar years, taken over by the Gestapo in 1933 and filled with data on people persecuted by the Gestapo in the government district of Osnabrück until 1945. At that time, 48.676 cards had been created with information on 48.015 individuals and 40.939 documented Gestapo actions against these individuals. Reliable data extraction methods allow retrieval of the victim's names and build a digital model of the card file index for analytical purposes.

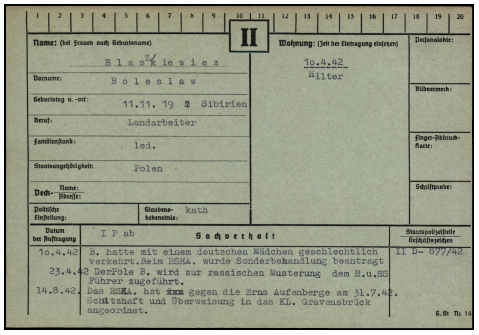

Fig. 1. An example of a semi-structured document image of the Gestapo dataset from the Lower Saxony State Archive (NLA OS, Rep 439, No. 3227).

III. Weakly Supervised Information Detection

The training process of traditional neural information extraction methods relies heavily on manually annotated data. Unfortunately, the annotation process is still an expensive and non-generalizing procedure. To overcome this limitation, we propose an efficient annotation-free information extraction approach for semi-structured document images, see Figure 2. Our approach assumes a collection of document images as an input that are derived from a series of templates. The overall concept is to create synthetic document images that are representative for the collection and use them to train an information extraction model. In order to create a realistic synthetic dataset we first extract the distinct templates from the collection. This is accomplished by transforming the document images into feature representations by resizing and flattening them to a fixed vector size. The representative templates are extracted by applying the DBSCAN clustering algorithm to these feature representations. Hereby, the corresponding document image of the obatined centroids are used. Once the templates are identified, the next step involves erasing the form content to generate cleaned template images. This process is accomplished by a manual procedure that utilizes the neural in-painting Lama model [1]. The user defines regions for removal through scribble annotations, taking only 1-2 seconds per image. The in-painting model reconstructs the annotated regions in a context-sensitive manner. To facilitate the synthesis pipeline, template regions are annotated,

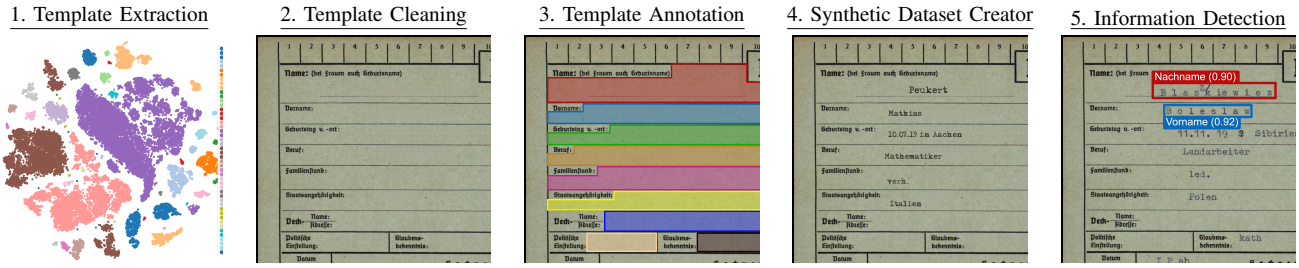

Fig. 2. An overview of our proposed information extraction pipeline for a collection of semi-structured document images. This framework consists of several steps to derive a synthetic dataset that is representative for the collection and to train an information detection model on these synthetic document images.

specifying the class of the region and the area where text can be located. Subsequently, a dataset of synthetic document images is created based on this information. For text detection, a YOLOv8 model [2] is fine-tuned on the synthetic dataset and used to extract bounding boxes on word image level with corresponding classes for relevant text information from document images. The final step involves a transcription of the cropped word images using a text recognition model.

IV. Self-Training

After the detection system extracts a document region that contains the relevant information with respect to the template annotations, a text recognition model is required to generate the respective transcription. In the application scenario annotated data is usually not available for training or evaluating text recognition models. Given the previously described steps of template extraction and cleaning, a potentially infinite number of synthetic text images can be generated. Training a neural network for text recognition only on synthetic data usually results in comparable poor performances. To improve recognition results without introducing further annotation demand, we follow a strategy known as self-training. In our experiments, we train and evaluate a sequence-to-sequence text recognition model [3] on several handwriting benchmark datasets without using any manually annotated data. Figure 3 depicts the general self-training approach. First, we train an initial recognition model solely on synthetic data. The model is used to make predictions for an unlabeled dataset. The predictions are then considered as pseudo-labels and training is continued on the pseudo-labeled data. The entire process of predicting and training on psuedo-labels is performed iteratively. To further improve performances, we apply a confidence-based selection scheme. The sigmoid activations of the network can be considered a confidence estimate and can be summarized into a quantitative measure on how likely a prediction is to be correct [4]. During self-training, pseudo-labels below a certain confidence threshold are neglected.

See Table I, for quantitative results on the four datasets of George Washington (GW), Bentham (BT), the IAM and CVL database. For more details on the respective benchmarks and the training setup, see [4]. We show that training on pseudo-labels improves performances compared to the purely synthetic baseline. Focusing on more confident predictions benefits

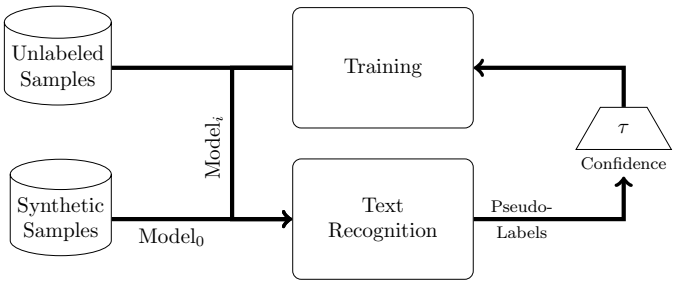

Fig. 3. Self-training for Text Recognition. First, an initial Model$_0$ is trained on synthetic data. Training is then performed on iteratively predicted pseudo-labels. A confidence measure is employed to threshold erroneous samples.

TABLE I
Experiments on different benchmarks from the literature

Selection	GW		IAM		CVL		BT	
	CER	WER	CER	WER	CER	WER	CER	WER
Synthetic	17.8	45.0	21.1	48.4	28.2	58.1	23.5	52.3
Self-Train (ST)	14.2	35.1	13.0	35.9	10.5	31.1	14.7	38.5
ST (Confidence)	**12.6**	**31.2**	**10.2**	**27.7**	**7.5**	**24.6**	**10.2**	**27.7**

performances on all benchmarks. Using network activations to quantify prediction quality constitutes only a rough confidence estimate.

V. Conclusion

Our work shows that it is feasible to establish a well performing information extraction pipeline in the absence of annotated training data. Following an annotation-free paradigm offers a potential solution to access the vast amount of information contained in big scale archive document collections.

References

[1] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Winter Conf. on Applications of Computer Vision*, 2022, pp. 2149–2159.

[2] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[3] L. Kang, J. I. Toledo, P. Riba, M. Villegas, A. Fornés, and M. Rusiñol, "Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition," in *German Conference on Pattern Recognition*, vol. 11269, Stuttgart, Germany, 2018, pp. 459–472.

[4] F. Wolf and G. A. Fink, "Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation," in *Proc. Int. Conf. on Pattern Recognition*. Montreal, Canada: IEEE, 2022, pp. 3885–3892.