Recognition-free Question Answering on Handwritten Document Collections

Oliver Tüselmann[0000−0002−8892−3306], Friedrich Müller[0000−0003−1885−6205], Fabian Wolf[0000−0001−8842−3718], 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. In recent years, considerable progress has been made in the research area of Question Answering (QA) on document images. Current QA approaches from the Document Image Analysis community are mainly focusing on machine-printed documents and perform rather limited on handwriting. This is mainly due to the reduced recognition performance on handwritten documents. To tackle this problem, we propose a recognition-free QA approach, especially designed for handwritten document image collections. We present a robust document retrieval method, as well as two QA models. Our approaches outperform the state-of-the-art recognition-free models on the challenging BenthamQA and HW-SQuAD datasets.

Keywords: Visual question answering · Information retrieval · Handwritten documents · Document understanding

1 Introduction

Question Answering (QA) is still an open and major research topic in a wide variety of disciplines [16,26,31]. Especially, the communities of Computer Vision (CV) and Natural Language Processing (NLP) focus on this task and made considerable progress [26,31]. Over the last few years, the Document Image Analysis (DA) community has shown an increasing interest in QA [15,16]. The majority of DA approaches tackle this task by adapting and using models from the NLP and CV communities [15,19,27]. Thereby, the text from a document image is transcribed and an answer is determined using a textual QA system [15,27]. This already leads to high performances for machine-printed document images with low recognition error rates [16]. However, the performances of these approaches decrease considerably on handwritten document images [15,16]. This is mainly due to the considerably reduced recognition accuracy, even though, substantial progress has been made in handwritten text recognition (HTR) over the last few years [8].

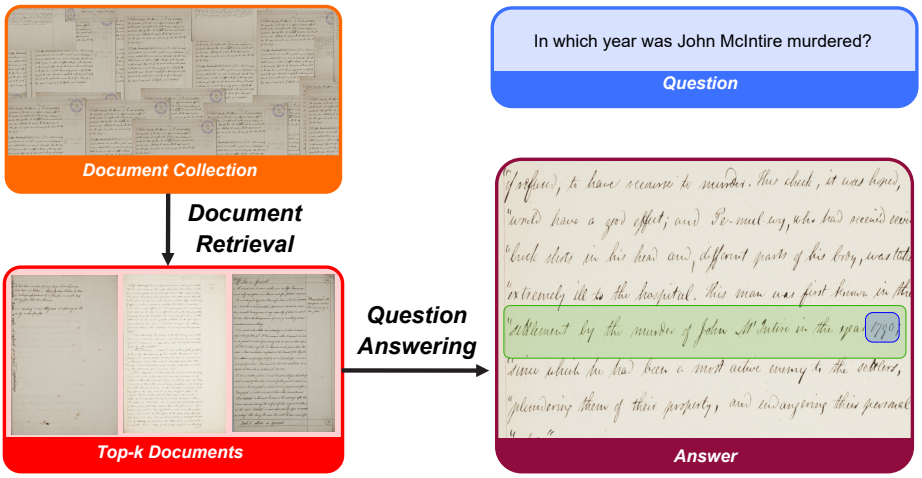


Fig. 1: An overview of the Question Answering pipeline on document image collections. Given a textual question and a document image collection, a document retriever identifies the k most relevant documents from the collection for answering the question. Finally, a word (blue) or line (green) region from one of these k document images is returned as the answer.

Answering questions on handwritten document images requires models that are robust with respect to handwriting recognition errors or do not rely on textual input. This is particularly important for QA on unknown document collections, as training data is usually not available and therefore, high handwriting recognition error rates are expected. For developing and evaluating such approaches, Mathew et al. recently proposed the BenthamQA and HW-SQuAD datasets [15]. These datasets provide questions as strings in natural language and expect answers as image regions of a document image rather than a textual response. Finding an answer in a single document image is already a challenging task. However, it is even more complicated in real world scenarios as a large collection of document images is often given. Therefore, it is crucial to identify documents in the collection that are relevant to answer the question and afterwards extract the answer from these documents. Fig. 1 provides an overview of this pipeline.

In this work, we propose a recognition-free approach for answering questions on handwritten document image collections. We present a robust document retriever as well as two QA approaches. The first QA model is based on the approach of Mathew et al. [15] and replaces their aggregation strategy with an attention based method. The second model is based on a QA architecture from the NLP domain and enables recognition-free QA on both word and line level. We compare our approach with recognition-free as well as recognition-based QA approaches on the challenging BenthamQA [15] and HW-SQuAD [15] datasets and are able to outperform state-of-the-art results by a large margin.

2 Related Work

QA on document collections usually requires a two-stage approach consisting of a document retriever and a QA model. We provide an overview on textual document retrieval (see section 2.1) and QA in the visual, textual as well as document image domain (see section 2.2).

2.1 Document Retrieval

Document retrieval is an information retrieval task that receives a textual request and returns a set of documents from a given document collection that best matches the query. Traditional approaches rely on counting statistics between query and document words [6,17]. Different weighting and normalization schemes over these counts lead to Term Frequency-Inverse Document Frequency (TF-IDF) models, which are still popular [6,17]. However, these models ignore the position of occurrences and the relationships with other terms in the document [6]. Therefore, models have been developed that can learn the relevance between questions and documents. Learning-To-Rank (LTR) [11] is a well-known document retrieval approach, which represents a query-document pair as a vector of hand-crafted features and trains a model to obtain similarity scores. Recently, deep neural ranking models outperform LTR models [18]. For a detailed overview on document retrieval, see [6,17].

2.2 Question Answering

QA is applied in various domains, leading to large variations among approaches. We present an overview on purely textual QA approaches as well as Visual Question Answering (VQA) models from the CV domain. Furthermore, we discuss recent progress for VQA on document images.

Textual Question Answering The textual QA community is mainly focusing on the Machine Reading Comprehension (MRC) [30] and OpenQA [31] tasks. In MRC, only one document is given and the answer is a snippet of the document. There is also an extension for this task, whereby models have to decide whether a question is answerable based on the document [30]. Traditional MRC approaches are mainly implemented based on handcrafted rules or statistical methods [30]. Long Short-Term Memory (LSTM)-based models with attention [22] achieved further progress in this field. In recent years, Transformer models (e.g. BERT [4]) improved the results considerably [22]. These models benefit from largely pretrained word embeddings, which encode useful semantic information between words. Currently, specialized transformer models (e.g. LUKE [28]) lead to state-of-the-art results. In contrast to MRC, OpenQA tries to answer a given question without any specified context. It usually requires the system to search for relevant documents in a large document collection and generate an answer based on the retrieved documents. OpenQA models are mainly a combination of document retrieval and MRC-based approaches [31]. For a detailed overview of textual QA, see [31].

Visual Question Answering Given an image and a query in natural language, Visual Question Answering (VQA) tries to answer the question using visual elements of the image and textual information from the query [26]. Most approaches rely on an encoder–decoder architecture, which embed questions and images in a common feature space [5,12,21]. This allows learning interactions and performing inference over the question and the image contents. Practically, image representations are obtained with Convolutional Neural Networks (CNNs) pre-trained on object recognition [12,26]. Text representations are obtained with word embeddings pre-trained on large text corpora. RNNs are used to handle the variable size of questions. Further progress in this field has been made using attention [2]. The attention mechanism allows the model to assign importance to features from specific regions of the image. Recently, Transformer based architectures achieved state-of-the-art results on multiple VQA benchmarking datasets [9]. For a detailed overview of VQA, see [26].

Document Image Visual Question Answering Mainly due to several new competitions [14,16] and datasets [14,15,16], there has been major progress in the area of answering questions on document images. These datasets provide MRC [14,16] as well as OpenQA tasks [15]. The approaches and datasets mainly focus on machine-printed documents, which contain visual and structural information (e.g. charts, diagrams) [15,19,27]. The layout is important for answering most of the questions [16,19,27]. The approaches are based on textual recognition results and adapt state-of-the-art QA systems from the NLP domain [19,27]. Recently, Mathew et al. [15] published a first dataset for QA on handwritten document collections. Furthermore, they proposed a recognition-free QA approach, which outperforms recognition-based QA models on handwritten datasets [15].

3 Method

In this section, we present our recognition-free approach for answering questions on document image collections. The approach consists of a document retriever (see section 3.2) and a QA model (see section 3.3). Both models are based on the robust Pyramidal Histogram of Characters (PHOC) attribute representation (see section 3.1). Given a query and a collection of word and line-segmented document images, our document retrieval approach assigns a score to each document image, indicating its relevance to answer the query. For each of the K most relevant documents, our QA model determines answer snippets and an associated confidence score. Finally, the snippet with the highest score is returned as the answer.

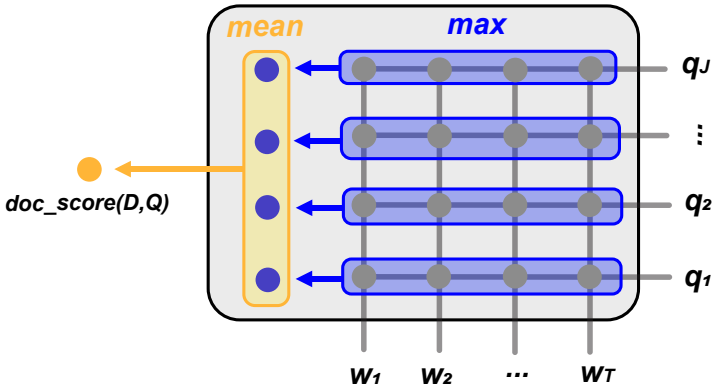


Fig. 2: Our attention based retrieval approach for calculating the similarity between a query $\mathbf{Q} = [\mathbf{q}_1, ..., \mathbf{q}_J]$ and a document or snippet $\mathbf{D} = [\mathbf{w}_1, ..., \mathbf{w}_T]$.

3.1 Query and document representation

To compute a similarity score between a document image and a query, as well as for question answering, the question words and the document images have to be transformed into a vector representation. Since we follow a recognition-free approach and the question is provided in a textual and the documents in a visual form, we use the Pyramidal Histogram of Characters (PHOC) representation that allows a robust mapping of words and images into the same space. A PHOC is a binary pyramidal representation of a character string and is used to represent visual attributes of a given word image. The embedding is successfully and widely used in the word spotting domain [1,10,23]. We use the TPP-PHOCNet [23] to realize a mapping from word images to a PHOC representation. The representations of the word images are finally stored in the order of their occurrences in the document image.

3.2 Retrieval

To determine the most relevant documents regarding a query in a given collection, we follow a similar approach as described in [15]. Hereby, they aggregate the question and the documents into vector representations of fixed size, by using the Fisher Vector framework [7]. Finally, the best matching documents are obtained by calculating the cosine similarity between the question and document embeddings. In contrast, our approach does not rely on such aggregation methods and instead uses the similarity between each word image from a document D and each question word from the pre-processed query Q as described in equation 1 and visualized in Fig. 2.

$$doc\_score(D, Q) = \frac{1}{|Q|} * \sum_{q \in Q} \max_{w \in D} [sim(w, q)] \tag{1}$$

We use the cosine similarity as the similarity measure. For each PHOC encoded question word $q \in Q$, the maximum similarity between q and the predicted PHOC vectors of the word images $w \in D$ is calculated. The overall similarity between Q and D is the averaged value over all these similarity scores and is computed for each document in the collection. Finally, the documents from the collection are sorted in descending order with respect to the calculated scores and the first K documents are returned as the result. In the following, we denote this approach as *Attention-Retriever*.

3.3 Question Answering

The recognition-free QA approach from [15] transforms document images into a set of two-line image snippets. For each of these snippets, an aggregated vector representation is determined based on the corresponding word images and is used to compute a similarity score with respect to an aggregated query vector. The score represents the confidence of finding the answer in the corresponding document region and determines the final answer of the system. Even though this approach can correctly locate the answers for some questions, the intuition behind this method is fairly questionable. The approach does not learn any real relationship between context and question, but exploits the heuristic that question words often occur close to the answer. Therefore, the approach does not realize a classical QA system, but rather an adapted syntactic word spotting approach for snippets.

NLP models are mainly based on the successes in transfer learning, where contextualized word embedding models were pre-trained on very large text collections. Unfortunately, transfer learning on handwritten word images is currently difficult and a robust mapping of word images into a semantic space is challenging even for static semantic word embeddings [24]. Therefore, it is currently not straightforward to adapt state-of-the-art NLP approaches to this task. However, there are previous state-of-the-art QA models from the NLP domain that do not rely on contextualized word embeddings and still lead to high performances on most datasets.

We follow the approach of the textual Bidirectional Attention Flow for Machine Comprehension (BIDAF) [22] model from the NLP domain and adapt it to a recognition-free QA model working on line instead of word level (see Fig. 3). The architecture can be divided into the word embedding, phrase embedding, attention flow, modeling, line embedding and output parts.

In the word embedding layer, all word images from a given document as well as the textual question words are represented as PHOC vectors. Here, the word images from the documents are transferred into the PHOC representation using the TPP-PHOCNet [23]. We further encode the line correspondence of each word image in the document using the positional encoding strategy from [4] and append them to the corresponding PHOC representations. The phrase embedding part uses a BLSTM to extract and model the temporal interactions

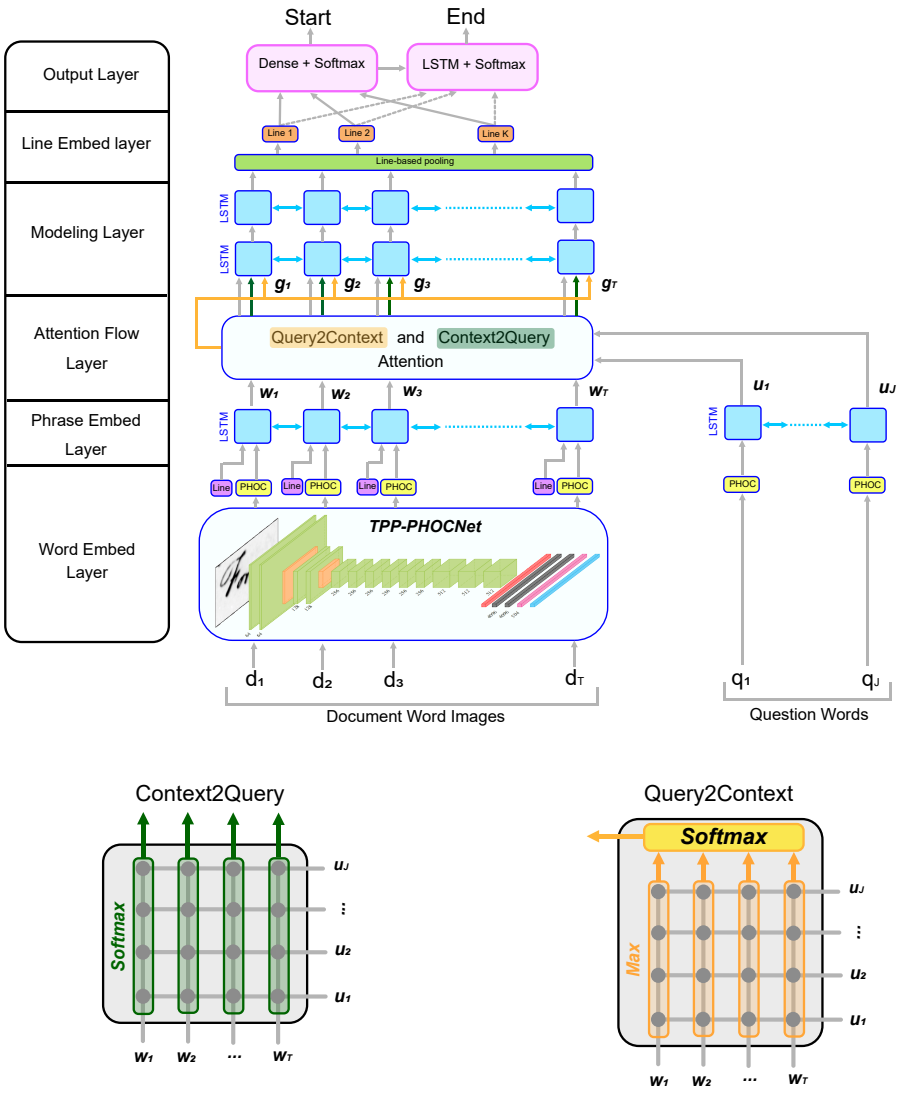


Fig. 3: The adapted BIDAF architecture for recognition-free Question Answering on line level.

between the word image representations from a given document as well as for the question word representations. Afterwards, the attention flow layer determines two types of attention scores between the obtained context $(\mathbf{w}_1, ..., \mathbf{w}_T)$ and question $(\mathbf{u}_1, ..., \mathbf{u}_J)$ vectors, namely Context2Query and Query2Context. Thereby, Context2Query signifies which context words have the closest similarity to one of the query words and are hence critical for answering the query and Query2Context signifies which query words are most relevant to each context word. Both of these attentions are based on a shared similarity matrix $\mathbf{S}$ between the contextual embeddings of the context $\mathbf{w}$ and the query $\mathbf{u}$. Hereby, $\mathbf{S}_{t,j}$ indicates the similarity between t-th context word and j-th query word and is computed by a trainable scalar function. The contextual embeddings $\mathbf{w}$ and the attention vectors are combined together to yield $\mathbf{g}$, where each vector can be considered as the query-aware representation of each context word. The obtained representations serve as the input to another two-stage BLSTM architecture, which models the relationship between questions and contexts. In this process, the BLSTM has as many outputs as the number of words in the document. The outputs of the BLSTM are reduced to the number of lines in the document by summing the word representations according to their line membership in the document. A dense layer is applied to each of these line representations and a softmax operation is performed. The result represents the pseudo-probability distribution for the start line of the answer. For calculating a similar distribution for the end line, the line representations are fed into another BLSTM and a dense layer as well as a softmax operation is applied to its output. The confidence for the prediction is the sum of the values before the softmax operation for the predicted start and end line indices.

The architecture can be used for word-level predictions by removing the line embedding layer. In the following, we will refer to the line-level model as *BIDAF-Line* and the word-level model as *BIDAF-Word*. In addition, we will refer to the adapted recognition-free QA approach of [15] as *Attention-QA*.

4 Experiments

We evaluate our proposed recognition-free QA approach on the HW-SQuAD and BenthamQA datasets (see section 4.1). Section 4.2 presents the implementation details and section 4.3 discusses the evaluation results. The performance of the QA systems is measured using the Double Inclusion Score (DIS) as introduced in [15] and shown in equation 2. The Small Box (SB) includes the word images that contain the answer. The Large Box (LB) includes all word images from those lines that are part of the SB as well as those from the lines above and below it. The Answer Box (AB) includes the word images from the lines predicted by the QA system. A visual example of these box definitions is given in Fig. 4. The prediction is considered a correct answer if the score is above 0.8.

$$DIS = \frac{AB \cap SB}{|SB|} \times \frac{AB \cap LB}{|AB|} \tag{2}$$

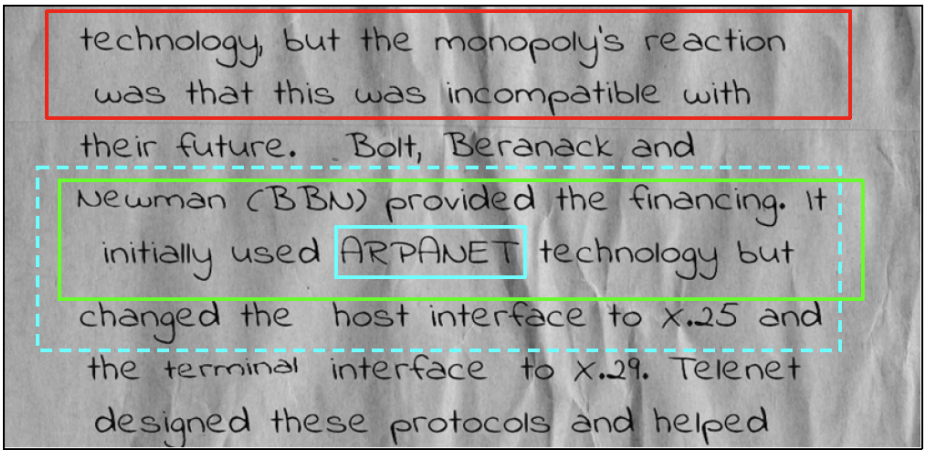


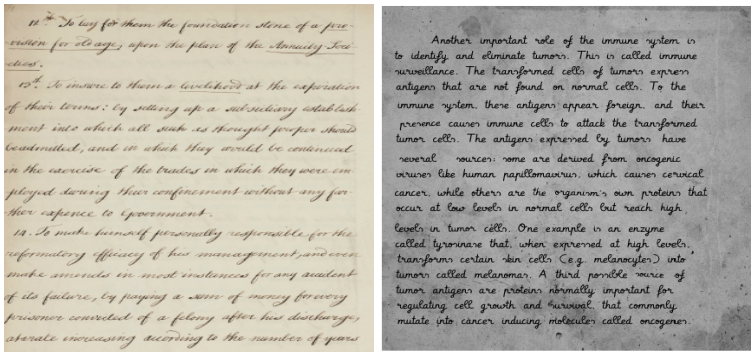
Fig. 4: An example of a correct (green) and an incorrect (red) predicted Answer Box for the question "Telnet used what interface technology?". Here, the rectangle around the word "ARPANET" represents the Small Box, whereas the dashed rectangle illustrates the Large Box.

4.1 Dataset

We train and evaluate our models on the recently proposed HW-SQuAD and BenthamQA datasets [15] (see Fig. 5), which contain question-answer pairs on handwritten document image collections in the English language. The datasets vary considerably in their size and characteristics and include synthetically generated as well as real handwritten documents.

BenthamQA [15] is a small historical handwritten QA dataset where questions and answers were created using crowdsourcing. The historic dataset contains 338 documents written by the English philosopher Jeremy Bentham and shows some considerable variations in writing styles. The dataset provides only a test set consisting of 200 question-answer pairs on 94 document images. The remaining 244 documents from the collection are used as distractors.

HW-SQuAD [15] is a QA dataset on syntactically generated handwritten document images from the textual SQuAD1.0 [20] dataset. The textual dataset is actually defined for an MRC task and was adapted by Mathew et al. [15] to an OpenQA task. The synthetic dataset consists of 20963 document pages containing a total of 84942 questions. The official partitioning splits the dataset into 17007 documents for training, 1889 for validation and 2067 for testing. Thereby, the training, validation and test sets contain 67887, 7578 and 9477 questions respectively.



(a) BenthamQA (b) HW-SQuAD

Fig. 5: Example documents for the BenthamQA and HW-SQuAD datasets.

4.2 Implementation Details

Our proposed document retriever relies on pre-segmented word images and our QA approaches also need line annotations. For our experiments, we use the gold-standard word and line bounding boxes available with the datasets. Questions are split into words and stopwords are removed using NLTK [3]. For training the BIDAF architecture, we use the HW-SQuAD dataset. We do not change the proposed parameters from [22] and use a hidden layer size of 100 as well as dropout with probability 0.2 for the BLSTMs. For optimization, we use ADADELTA [29] with the Cross Entropy loss and a learning rate of 0.5. The positional line encoding produces a 30-dimensional vector using sine and cosine functions.

For word representation, we use a 504-dimensional PHOC vector consisting of lowercase letters (a-z), numbers (0-9) and the levels 2, 3, 4 and 5. We pre-train the TPP-PHOCNet on the HW-SQuAD [15] as well as IIIT-HWS [10] datasets and fine-tune the model on the IAM database [13]. We use a batch size of 40 and a momentum of 0.9. The parameters of the network are updated using the Stochastic Gradient Decent optimizer and the Cosine loss. The learning rate is set to 0.01 during pre-training and 0.001 while fine-tuning. It is divided by two if the loss has not decreased in the last three epochs. We binarize the word images to remove the background from text. This is especially important as the background of document images from BenthamQA largely differ from those in IIIT-HWS and IAM.

4.3 Results

In this section, we show the evaluation performances of our recognition-free QA approach on handwritten document image collections. We evaluate and compare the document retrieval approach in section 4.3 and the three QA approaches in section 4.3. Finally, we present and discuss the results on the combination of the retrieval and QA approaches in section 4.3. For all subsequent experiments, we use the Attention-Retriever approach presented in section 3.2.

Table 1: Top-5 accuracy (%) for document retrieval approaches.

	Approach	HW-SQuAD	BenthamQA
Pred.	Mathew et al. (rec-free) [15]	46.5	55.5
	Mathew et al. (rec-based) [15]	86.1	32.0
	Attention-Retriever	**86.2**	**92.5**
GT	Mathew et al. (rec-based) [15]	90.2	98.5
	Attention-Retriever	87.2	98.0

Retrieval For answering a given question in a document collection, it is common to determine the five most relevant documents with respect to the query. To evaluate those retrieval models, we use the Top-5 accuracy as described in [15]. This score represents the percentage of questions from a given test set that have their associated answer document in the top five predicted retrieval results. In table 1 we present the Top-5 accuracy scores for our document retrieval approach and compare it to the literature. We also show the results of NLP models working on ground truth transcriptions in this table.

The results show that we are able to improve the state-of-the-art Top-5 accuracy scores on both datasets. We clearly outperform the recognition-free approach from the literature. On the HW-SQuAD dataset, we perform marginally better compared to the recognition-based approach proposed in [15]. The recognition based model performs on nearly perfect recognition results (97.9% word accuracy). When the recognition performance becomes worse as in BenthamQA (23.2% word accuracy), the vulnerability of the approach to recognition errors is revealed and only a low performance can be achieved. The results from our model with ground truth and predicted PHOCs are almost identical, demonstrating the robustness of our approach. The differences between HW-SQuAD and BenthamQA can be explained by the lower prediction performance of the PHOC vectors for BenthamQA. In this case, the TPP-PHOCNet can achieve a query-by-string score of 98.5 on HW-SQuAD and 77.3 on BenthamQA. Interestingly, the NLP method working on ground truth transcriptions can only achieve marginally higher scores compared to our attention approach, demonstrating the capabilities of our model.

Question Answering In order to evaluate the performance of our three proposed QA approaches without the influence of the retrieval model, we evaluate their performances on a MRC rather than the OpenQA task. Thus, the QA systems only work on the document that contains the answer to the question. Table 2 shows the results of our QA approaches as well as upper bounds using a state-of-the-art QA approach (BERT [4]) working on ground truth transcriptions.

Table 2: Machine Reading Comprehension. Performance measured in DIS.

	QA-Approach	HW-SQuAD	BenthamQA
Pred.	Attention-QA	47.5	38.5
	BIDAF-Word	57.2	28.0
	BIDAF-Line	**68.1**	**50.5**
GT	Attention-QA	47.7	39.0
	BIDAF-Word	57.7	47.0
	BIDAF-Line	68.7	62.0
	BERT [4]	94.4	88.0

The results show that our line-based BIDAF model can achieve higher scores compared to the word-based model and the attention based approach. A comparison with the literature is not possible, as Mathew et al. do not evaluate their approaches on this task. As already shown for the document retrieval, a similar relationship emerges between the performances of the models based on ground truth and predicted PHOCs. However, compared to the attention approach, the PHOC prediction errors have a stronger impact on the performances of the BIDAF models. In comparison to the line-based approach, the word-based BIDAF model seems to be quite sensitive to erroneous PHOC predictions. Presumably, the line-based model is less sensitive to the recognition errors due to the aggregation on line level. It should be noted, that the performances of our approaches show potential for improvement compared to NLP models working on textual annotations. This gap is likely due to the successful application of transfer learning in the textual domain.

End-to-End Question Answering In the previous subsections, we have evaluated the individual components of our system. For answering questions in document collections, a combination of those is required. Table 3 shows the results for the combination of our document retriever and the line-based BIDAF model as well as approaches from the literature. For this evaluation, the document retrieval approaches extract the top five documents from the collections.

Our approach can clearly outperform the recognition-free method proposed by Mathew et al. [15]. The recognition-based system from [15], however, outperforms our approach on the HW-SQuAD dataset, but clearly fails on BenthamQA. This supports the common research outcomes, whereas the performances of textual NLP models are quite high on datasets with low recognition errors, but decrease considerably when the amount of recognition errors rise [25]. The results show that the PHOC prediction errors affect the performance of our model, however, it shows a considerably improved robustness compared to recognition-based models.

Table 3: End-to-end answer line snippet extraction. Performance measured in DIS.

	QA-Approach	HW-SQuAD	BenthamQA
Pred.	Mathew et al. (rec-free) [15]	15.9	17.5
	Mathew et al. (rec-based) [15]	**59.3**	2.5
	BIDAF-Line	45.0	**37.5**
GT	Mathew et al. (rec-based) [15]	74.8	74.0
	BIDAF-Line	45.3	55.0

5 Conclusions

In this work, we present a recognition-free question answering system for handwritten document image collections. The system consists of an attention based document retriever as well as a question answering approach. Our document retrieval model achieves new state-of-the-art scores for the retrieval task on all considered datasets. For question answering, textual approaches benefit from transfer learning methods and outperform recognition-free approaches on the HW-SQuAD dataset with low word error rates. Considering the desired application to historical datasets with presumably no annotated training material, error rates are usually significantly higher. As seen on BenthamQA, this leads to a considerable decrease of the QA performance for recognition-based models. Our experiments show the robustness of our proposed combination of recognition-free retrieval and QA system and that it is able to outperform recognition-free as well as recognition-based methods from the literature.

References

1. 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)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Int. Conf. on Computer Vision and Pattern Recognition. pp. 6077–6086. Salt Lake City, UT, USA (2018)
3. Bird, S.: NLTK: The natural language toolkit. In: Annual Meeting of the Association for Computational Linguistics. pp. 69–72. Sydney, Australia (2006)
4. 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 - Human Language Technologies. pp. 4171–4186. Minneapolis, MN, USA (2019)
5. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proc. Conf. on Empirical Methods in Natural Language Processing. pp. 457–468 (2016)

6. Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W.B., Cheng, X.: A deep look into neural ranking models for information retrieval. Information Processing and Management **57**(6), 102067 (2020)
7. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Int. Conf. on Neural Information Processing Systems. pp. 487–493. Denver, CO, USA (1998)
8. 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: Proc. German Conf. on Pattern Recognition. pp. 459–472. Stuttgart, Germany (2018)
9. Khan, A.U., Mazaheri, A., da Vitoria Lobo, N., Shah, M.: MMFT-BERT: Multimodal fusion transformer with BERT encodings for visual question answering. In: Proc. Conf. on Empirical Methods in Natural Language Processing. pp. 4648–4660. Online (2020)
10. 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)
11. Liu, T.: Learning to rank for information retrieval. Found. Trends Inf. Retr. **3**(3), 225–331 (2009)
12. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Int. Conf. on Computer Vision. pp. 1–9. Santiago, Chile (2015)
13. 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)
14. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa. CoRR **abs/2104.12756** (2021)
15. 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)
16. Mathew, M., Karatzas, D., Jawahar, C.V.: DocVQA: A dataset for VQA on document images. In: IEEE Workshop on Applications of Computer Vision. pp. 2199–2208. Waikoloa, HI, USA (2021)
17. Mitra, B., Craswell, N.: An introduction to neural information retrieval. Foundations and Trends in Information Retrieval **13**(1), 1–126 (2018)
18. Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., Cheng, X.: DeepRank: A new deep architecture for relevance ranking in information retrieval. In: Proc. ACM Int. Conf. on Information and Knowledge Management. pp. 257–266. Singapore (2017)
19. Powalski, R., Borchmann, L., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Palka, G.: Going Full-TILT boogie on document understanding with text-image-layout transformer. In: Proc. Int. Conf. on Document Analysis and Recognition. pp. 732–747. Lausanne, Switzerland (2021)
20. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100, 000+ questions for machine comprehension of text. In: Proc. Conf. on Empirical Methods in Natural Language Processing. pp. 2383–2392. Austin, TX, USA (2016)
21. Saito, K., Shin, A., Ushiku, Y., Harada, T.: DualNet: Domain-invariant network for visual question answering. In: Int. Conf. on Multimedia and Expo. pp. 829–834. Ypsilanti, MI, USA (2017)
22. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. In: Int. Conf. on Learning Representations. Toulon, France (2017)

23. 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 (2016)
24. 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)
25. 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)
26. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A.R., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163**, 21–40 (2017)
27. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florêncio, D.A.F., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: Annual Meeting of the Association for Computational Linguistics and Int. Joint Conf. on Natural Language Processing. pp. 2579–2591. Bangkok, Thailand (2021)
28. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep contextualized entity representations with entity-aware self-attention. In: Proc. Conf. on Empirical Methods in Natural Language Processing. pp. 6442–6454. Online (2020)
29. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. CoRR **abs/1212.5701** (2012)
30. Zeng, C., Li, S., Li, Q., Hu, J., Hu, J.: A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets. CoRR **abs/2006.11880** (2020)
31. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.: Retrieving and reading: A comprehensive survey on open-domain question answering. CoRR **abs/2101.00774** (2021)