Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 29–36
DOI: 10.15439/2024F7400
ISSN 2300-5963 ACSIS, Vol. 41

Enhancing Text Recognition of Damaged Documents through Synergistic OCR and Large Language Models

Thomas Asselborn*, Jens Dörpinghaus‡§, Faraz Kausar†, Ralf Möller*, Sylvia Melzer†
* Universität Hamburg, Institute for Humanities-Centered AI, Warburgstraße 28, 20354 Hamburg, Germany,
Email: {thomas.asselborn, ralf.moeller}@uni-hamburg.de
, https://orcid.org/0009-0005-3011-7626, https://orcid.org/0000-0002-1174-3323
† Universität Hamburg, Centre for the Study of Manuscript Cultures,
Warburgstraße 26, 20354 Hamburg, Germany,
Email: faraz.kausar@studium.uni-hamburg.de, sylvia.melzer@uni-hamburg.de, https://orcid.org/0000-0002-0144-5429
‡ Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
§ University of Koblenz, Germany,
Email: jens.doerpinghaus@bibb.de, https://orcid.org/0000-0003-0245-7752

***Abstract*—Optical Character Recognition (OCR) remains a highly relevant area of research in pattern recognition. Its applications span various domains, including supporting reading for the visually impaired, interpreting Morse codes, capturing postal addresses, evaluating emails, scanning price tags and passports, and extracting text from digitised documents. As the volume of digitised data continues to grow, challenges arise in capturing the semantic structure of documents through logical structure analysis and providing data suitable for information retrieval to answer specific research questions. While classic OCR processes like Tesseract and OCRopus work well for contemporary digitised documents, there is room for improvement in text and word recognition of historical documents that are severely damaged. Large Language Models (LLMs) like GPT-4 can be effectively used for text recognition tasks, utilising their advanced natural language processing capabilities to interpret and reconstruct unclear or damaged text, offering potential for improving the overall text recognition process. However, challenges arise additionally when documents contain e. g. a mixture of single-column and double-column text, images and text, or words not known or blocked by the agents.**

This article aims to find a suitable combination of OCR models and LLMs to accurately add missing words to texts according to their original versions.

I. Introduction

EVEN after 90 years, OCR (**O**ptical **C**haracter **R**ecognition) is still a very topical area of research in the field of pattern recognition. The areas of application for this technology are very wide-ranging. These include supporting reading for the blind, interpreting Morse codes, automatically capturing postal addresses, evaluating e-mails, scanning price tags and passports and extracting text from digitised documents [4]. The number of digitised data is steadily increasing and various repositories [17], [12], [32] with different search strategies have already been created. With regard to the data analysis of digitised data, there is an increasing challenge to capture the semantic structure of a document through logical structure analysis and to provide data that is suitable for retrieving information to answer individual research questions and consists of more than just a comparison of character strings. In addition, to promote document understanding, relevant information is required that not only results from text extraction, but also uses other data from images, notes, drawings, fonts, font colours, locations, document structures, etc.

The classic OCR processes include the open source OCR Tesseract or OCRopus. These methods have been extended with regards to the possibilities of machine learning to meet the above-mentioned challenges among others. One extension, for example, are the OCR-D software modules [23]. While the text recognition of these processes works very well for digitised documents from the present time, there is still room for improvement in the text and word recognition of historical documents that are severely damaged, e.g. by water damage, glued-on notes, perforations, holes, mould, etc. which can lead to words that are no longer recognisable. Some documents may be inscribed or obviously crossed out in later years, particularly in passages reflecting past ideology.

LLM (**L**arge **L**anguage **M**odel)s such as GPT-3 can be used effectively for text recognition tasks. Their advanced natural language processing capabilities allow them to interpret and reconstruct unclear text in addition to filling in spaces with damaged or missing text. These capabilities offer potential to improve the text recognition process. A simple implementation of first recognising text and then using an agent such as

The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796.

ChatGPT[1], Perplexity.ai[2] or UHHGPT[3] to add words becomes a challenge when documents contain a mixture of single-column and double-column text or images and text, or the words used are not part of the LLM (either not known or blocked by the agent).

Some OCR approaches were tested and it became apparent that text recognition of tables poses a major challenge. Only by correctly identifying the image and text regions that belong together can the text be generated in the correct order or the use of agents make sense. Prompt engineering also plays an important role in ensuring that an existing text is produced as it might be written, rather than a new text. The agent must first assume the role of the writer from the relevant time so that the desired correct text is produced. The challenge we face in this article is to find a suitable combination of OCR models or tools and an LLM so that the missing words from texts are added according to the original version of the text.

We have tested our approach on the digitised journals of the Godeffroy Natural History Museum[4], which existed in Hamburg from 1861 to 1885 and on a German legal document corpus. The Federal Institute for Vocational Education and Training maintains a collection of occupation-related documents with legal bases, which reflect about 85 years of German VET (**V**ocational **E**ducation and **T**raining) history. In recent years, this collection has been systematically recorded for the first time, resulting in precise knowledge of its contents on the one hand and the state of preservation of the individual documents on the other.

The results obtained provide an overview of various combined approaches and show that even with poor OCR results, the use of LLMs still delivers good results overall.

II. Related Work

Vocational Education and Training

Understanding the practicalities of reform implementation is crucial for effective VET policy transfer and adaptation.

A lot of different research has been done in the digitisation of documents in recent years: For example, historical Finnish newspapers, see [17], or historical publications of the Bundesanzeiger, see [12]. The historical development of vocational training regulations has only been studied to a very limited extent [15], while the general history and development of the labour market in relation to occupations receives much attention, see [37], [13], [29], [21]. Other works focus on the current development of regulations, see [18], and their analysis is also widely considered [10], [26], [2]. However, it remains unclear whether this is due to the fact that historical resources are currently not publicly available.

The Historical international standard classification of occupations (HISCO) is a publicly available dataset of comparable occupations that would be a prerequisite to make historical occupations and regulations interoperable. It was introduced in 2002 [19] and is available as a database at https://historyofwork.iisg.nl/index.php, where several datasets can be downloaded. However, the list of German occupations is incomplete. Another relevant dataset is prepared as Ontologie historischer, deutschsprachiger Berufs- und Amtsbezeichnungen (see https://www.geschichte.uni-halle.de/struktur/hist-data/ontologie/), but is currently not publicly available. Classifications for GDR (**G**erman **D**emocratic and **R**epublic) occupations are also not yet digitally available, while their mapping to standards like KldB is widely discussed [9], [1]. Another dataset is offered as "Genealogie der Berufe", but is only available as a web service (see https://www.bibb.de/dienst/berufesuche/de/index_berufesuche.php/). Also worth mentioning is the seminal work by Wolf-Dieter Gewande, who in 1999 for the first time compiled unpublished recognition data and traced the development of more than 1300 occupations to the present, see [11].

While very little research has been done on the historical regulation of vocational education and training in Germany, we can identify a second research gap: Data integration should be accompanied by linked data sets for occupational classifications that are not currently available. Thus, the integration of older data such as the KldB 1975, 1988 and 1992 is crucial.

Text recognition

To support as many application areas as possible with OCR, individual OCR modules were developed in various research projects, including the OCR-D project [23]. During the OCR-D project, the OCR-D software was developed to allow the easy combination of a variety of so-called processors – independent tools for specific tasks – to define workflows tailored to the peculiarities of different templates and thus automate the process of a large quantity of prints, particularly from the 16th to 18th centuries (cf. [33], [34], [35]. It is only through such automation that it is possible to make large collections available in their entirety as full-text to the scientific community.

OCR4all is a web application that offers a semi-automatic workflow tailored to digitise historical documents [30]. However, when compared to Tesseract, OCR4all's performance is not as strong. While OCRopus is an open source software, it also falls short of Tesseract's performance. ABBYY FineReader, a commercial tool, typically provides only slightly better results than Tesseract, see [14], [5].

LLMs for Text Recognition

LLMs such as GPT-3 and GPT-4 have shown remarkable capabilities in understanding and generating human-like text. Although they are primarily used for NLP (**N**atural **L**anguage **P**rocessing), recent research has investigated the use of LLMs to improve text recognition. [31], [36]

To the best of our knowledge none of the literature we found specifically discussing the synergies between LLMs and OCR techniques. This therefore appears to be a relatively new area of research with little direct literature available to date. The

[1]https://chat.openai.com
[2]https://www.perplexity.ai/
[3]https://uhhgpt.uni-hamburg.de/
[4]https://www.biodiversitylibrary.org/item/244246

sources indicate some related work on the use of LLMs for text recognition tasks, but a comprehensive overview of their synergies with OCR is not readily available based on these search results.

GPT4-o has demonstrated proficiency in multilingual applications. For German text input specifically, GPT4-o emerges as the preferred choice. Its superior performance in handling multilingual tasks, including those involving German, makes it particularly well-suited to process German text. [20]

III. Use Cases

Document collections constitute a vital component of vocational training research. VET and CVET (**C**ontinuing **V**ocational **E**ducation and **T**raining) in Germany have been subject to regulation since the 1920s. Over the course of many decades and through various political regimes, including the Third Reich, the German Democratic Republic, and the Federal Republic of Germany, these regulations have undergone significant evolution. The job archive at the Federal Institute for Vocational Education and Training (Bundesinstitut für Berufsbildung, BIBB) houses numerous historical VET and CVET regulations, which are largely inaccessible.

The key stakeholders in continuing vocational training in Germany include: (a) educational institutions, (b) companies and enterprises, (c) employees, and (d) sponsors. In light of the aforementioned transformation processes, it is imperative that these stakeholders adapt to the evolving conditions and requirements that they face. These challenges have shifted over time, necessitating a comprehensive overview and analysis of how regulatory documents reflect the aforementioned evolving requirements, changes, and challenges. For instance, it is crucial to identify which educational content is increasingly offered and demanded, in order to draw conclusions about the development needs of both the vocational and continuing education systems. The research-based development of the vocational education system aims not only to ensure the economy's competitiveness at a systemic level but also to combat unemployment and stabilise the social security system [8]. However, the historical regulations are not currently available in a digitised format. Given that these documents span a long period and multiple states (the German Empire, the GDR, and the FRG (**F**ederal **R**epublic of **G**ermany)), the challenges for OCR and data infrastructure are substantial.

The German Committee for Technical Education (DATSCH) was established in 1908, and from that point forward, a series of documents were created with the objective of standardising occupations. The age of the documents has resulted in the deterioration of the paper, with approximately two-thirds written in Fraktur script and the remainder in various Latin fonts. Preservation varies, with some documents exhibiting well-preserved characteristics and others displaying signs of water damage, glued notes, perforations, or mould. Some documents have later inscriptions or are crossed out, particularly passages reflecting Nazi ideology.

The majority of early order specifications are in DIN A5 format, with special formats ranging from pocket-sized job descriptions (DIN A6) to large inserts up to about DIN A1. A smaller part of the collection includes legal regulations from the Federal Republic of Germany, subject to BBiG, HwO, specific health profession laws, or federal school legislation. This collection encompasses training regulations, amendments, corrections, framework curricula, and advanced training regulations from the federal government, federal states, and competent bodies.

The GDR materials pertain to training and advanced education, including training documents for skilled worker training and socialist vocational training (Ausbildungsunterlagen für die Facharbeiterausbildung and Ausbildungsunterlagen für die sozialistische Berufsbildung), training plans (Ausbildungspläne), and equipment norms. The GDR documents are frequently bound as booklets or books, comprising up to 323 pages in A4 and A5 formats. They are typically printed in two columns with a typewriter font.

The diverse formats and conditions of the documents present significant OCR challenges. The use of Fraktur and multiple Latin scripts, along with physical damages complicates digitisation. Moreover, the prevalence of special formats, two-column layouts, and typewriter fonts in GDR materials serves to compound the difficulty in creating accurate digital copies. Consequently, novel approaches to support OCR on these very specific documents are of great importance for further research on (vocational) education.

At the Centre for the Study of Manuscript Cultures (CSMC) at the Universität Hamburg, there are a number of application areas in which OCR, like OCR-D, is used in a wide variety of workflows which can then support the evaluation of historical prints and documents. In the special research project "Sonderforschungsbereich" (SFB) 950: Manuscript Cultures in Asia, Africa and Europe, the empirical diversity of manuscript cultures was researched on the basis of the material. This resulted in numerous digital copies. Further digital copies have been and are being produced by the current DFG Cluster of Excellence 2176 - Understanding Written Artefacts (UWA). Automated character and word recognition of the digitised material using OCR can support research activities in the evaluation of historical prints and manuscripts. Historical manuscripts are often not in best condition and may include damages in a similar way described previously; potentially it may be even more dramatic. The application would not only save time, but it would also be conceivable to establish separate software modules to be included automatically after the standard OCR process to create a new, improved application so that damaged texts can also be restored.

IV. Optical Character Recognition (OCR)

OCR is the process of recognising text in a scanned or photographed document, image-only PDF containing text and similar types of documents. The goal of this process is to convert this text into a machine readable format, e.g. as a plain text file, so that further processing can be performed. [16] This process contains multiple discrete steps that are sometimes also called activities.

A. Activities

OCR modules are available and described in more detail in [24]. Some activities are presented in the following:

- *Binarization* is the process of converting an image into a binary representation, where each pixel is either black or white.
- *Dewarping* is used to correct distortions in scanned documents caused by the curvature of the page or the scanning process itself.
- *Despeckling* aims to remove small, isolated spots or noise from scanned images, improving their overall quality.
- *Deskewing* is the process of straightening a skewed or rotated document image, aligning it properly for further processing.
- *Font identification* analyses the shapes and characteristics of characters in a document to determine the font or fonts used.
- *Segmentation* is the process of dividing an image into meaningful regions or components for further analysis.
- *Region segmentation* identifies and separates different regions within a document, such as text blocks, images, and tables.
- *Region classification* categories the segmented regions of a document into different types, such as text, graphics, or tables.
- *Line segmentation* breaks down text regions into individual lines, enabling line-by-line processing.
- *Line recognition* analyses and interprets the content of each line, extracting relevant information.
- *OCR* is the process of converting scanned texts images into machine-readable text data.
- *Text recognition* analyses and interprets the segmented text, converting it into machine-readable format.

The existing OCR-D modules can be easily combined in an individual workflow, which ensures very good adaptability of the OCR-D modules, at least for the evaluation of digital prints.

B. OCR-D Workflow Application

As mentioned briefly above, the use of the OCR-D software modules and workflows for the full-text digitisation of the "Journal of the Godeffroy Museum". This collection was used because the results can be published without any problems in terms of copyrights, etc. and have the same representative requirements for text recognition of table contents.

Fig. 1 shows the result after running the Tesseract OCR module for a table. Green indicates which text characters were recognised correctly and red indicates which text characters were recognised incorrectly. We used the workflow:

1) Binarization
2) Region segmentation
3) Line segmentation
4) Line recognition

The workflow with just a few steps shows that an improvement in region recognition is necessary. Extending the workflow with additional steps usually leads to better results, but unfortunately this does not apply in the case of tables. A subsequent application of agents could lead to no meaningful texts, since row-by-row evaluation is the wrong reading direction for tables. Thus, for the recognition of texts, we have used other OCR tools.

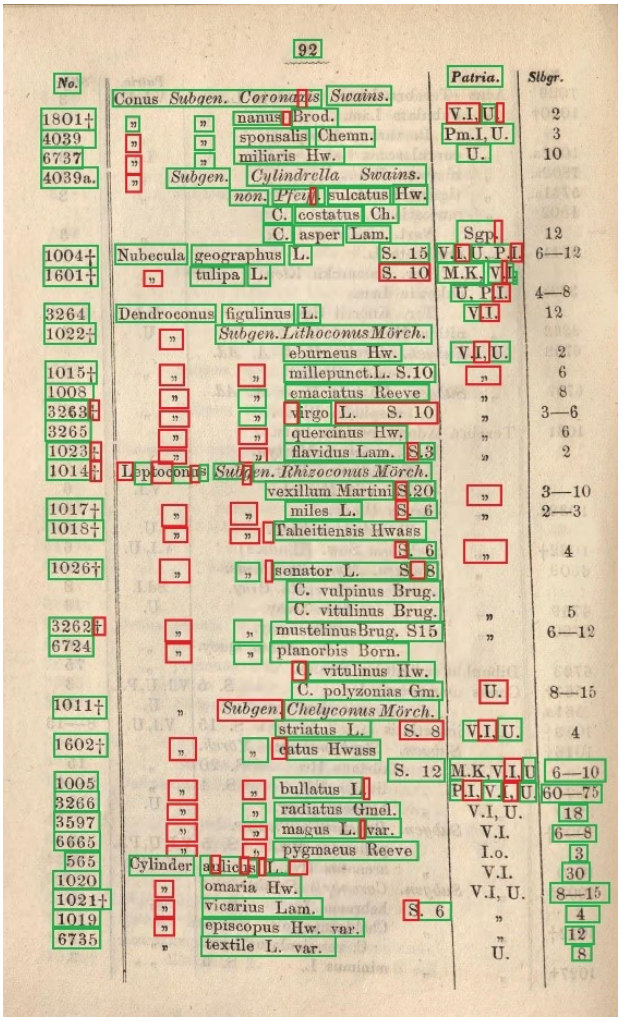

Fig. 1. Table from https://www.biodiversitylibrary.org/item/ 244246 page 92 with the OCR-D results

C. gImageReader

We also tested the OCR software gImageReader, which offers the possibility of correction by a user via a user interface. It is a GUI (**G**raphical **U**ser **I**nterface) front-end for Tesseract [22]. Thus, it provides the option for people that have no skills in using the command line to use Tesseract. Fig. 2 shows the areas marked by the software. The user has the option of adding further areas manually or deleting other areas. Here, the contiguous text areas of the table were not recognised convincingly well.

Using gImageReader on the example [3] from the VET corpus shown in Fig. 3, we got the following result:

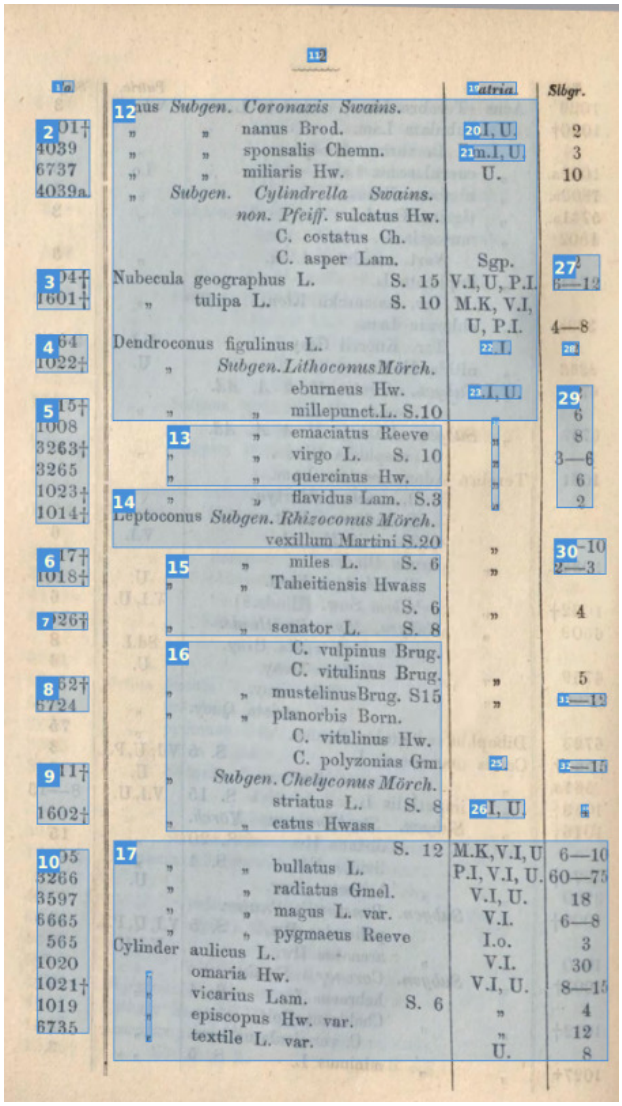

Fig. 2. The text areas marked by the OCR software gImageReader after performing the segmentation

```
1 Kenntnisse Uber die Aufgaben
inhalte gem. und Gliederung des Betriebes
Berufsbild und seine Einordnung in die
Gesamtwirtschaft

1.1 Art, Rechtsform und Gliederung des
Kenntnisse und Ausbildungsbetriebes
    beschreiben

-|1.9 Aufgaben der einzelnen Abteilungen

und ihre Zusammenarbeit erkldren

und Arbeitsabldufe im ausbilden-
den Betrieb beschreiben
```

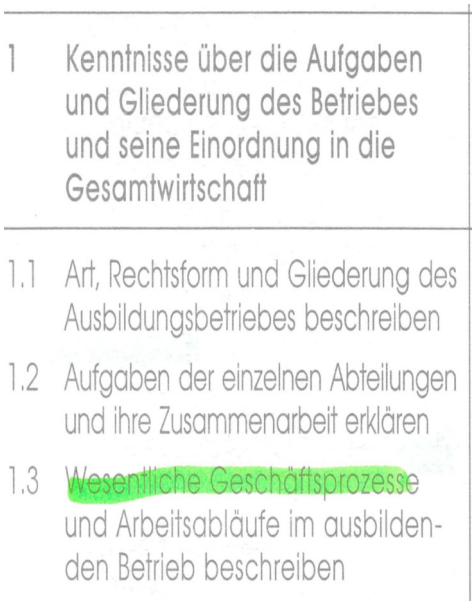

Fig. 3. Excerpt from "AKTUALISIERTE AUSBILDUNGSPLANEMPFEHLUNGEN Datenverarbeitungskaufmann Datenverarbeitungskauffrau" including human annotation [3]

D. OCR4all

We used OCR4all with the standard workflow as described in [25] to recognise the same text example shown in Fig. 3. Again, the results provided were not sufficient:

```
Kenntnisse über die Aufgoben
und Gliederung des Betriebes
und seine Einordnung in die
Gesomtwirtschoft
Ausbildungs-
holbjohre
afele
P
-
```

E. ABBYY FineReader

Finally, we used ABBYY FineReader 15 OCR-Editor to recognise the texts as well. Here, the result was better than what we got from all other OCR engines used in our sample testing.

The ABBYY FineReader output is:

```
1 Kenntnisse über die Aufgaben und Gliederung
    des Betriebes und seine Einordnung in die
    Gesamtwirtschaft
1.1 Art Rechtsform und Gliederung des
    Ausbildungsbetriebes beschreiben
1.2 Aufgaben der einzelnen Abteilungen und
    ihre Zusammenarbeit erklären
1 id V'V: vJtQ tz 1 rll r G 1 1 kZz \37\Zz0Gy
    | l-G-M |ö l övG
und Arbeitsabläufe im ausbildenden Betrieb
    beschreiben
```

Text recognition of tables works quite well with this tool. The texts that were marked in the original were not recognised. A jumble of letters was output. No meaningful search can be activated for this bad case. This can be remedied to a certain extend by a LLM, as shown next.

V. Combination of OCR and LLMs

ChatGPT is an advanced language model developed by OpenAI, designed to generate human-like text and engage in natural and coherent conversations with users. It is built upon the GPT (**G**enerative **P**re-trained **T**ransformer)-3.5 architecture. It can maintain context in conversations and respond in a contextually relevant manner, making it suitable for a wide range of applications. There are other agents, for example UHHGPT, that have been developed and use some of the same functions such as GPT-3.5 or GPT-4 omni. UHHGPT is a graphical user interface based on HAWK-Ki[5] that utilises the OpenAI API to send queries to and receive answers from one of the OpenAI GPT models. Its main goal is to provide anonymous access to the models since the user´s query cannot be traced back to the individual member of the university. Additionally, it offers access to GPT-4 omni for free for all members of the university. The prompt sent to OpenAI is always augmented by "You are a helpful assistant who works at the University of Hamburg in Germany." We have used UHHGPT with the function GPT-4 omni for our experiments. The following outputs show the same text excerpts from Section IV.

Prompt:

```
The following text was recognized after an
    OCR. Complete the words as they could
    have been written: [OCR output above is
    part of this prompt].
```

The output of UHHGPT with the function GPT-4 omni (OCR4all text) is:

```
Kenntnisse über die Aufgaben
und Gliederung des Betriebes
und seine Einordnung in die
Gesamtwirtschaft
Ausbildungsjahre
Anleitung
Praktikum
```

The output of UHHGPT with the function GPT-4 omni (gImageReader) is:

[5]https://github.com/HAWK-Digital-Environments/HAWKI

```
1. Kenntnisse über die Aufgaben
inhalte gemäß und Gliederung des Betriebes
    ,
Berufsbild und seine Einordnung in die
Gesamtwirtschaft.

1.1 Art, Rechtsform und Gliederung des
Kenntnisse- und Ausbildungsbetriebes
    beschreiben.

1.9 Aufgaben der einzelnen Abteilungen
und ihre Zusammenarbeit erklären
und Arbeitsabläufe im ausbildenden
Betrieb beschreiben.
```

The output of UHHGPT with the function GPT-4 omni (ABBYY FineReader text) is:

```
1 Kenntnisse über die Aufgaben und
    Gliederung des Betriebes und seine
    Einordnung in die Gesamtwirtschaft
1.1 Art, Rechtsform und Gliederung des
    Ausbildungsbetriebes beschreiben
1.2 Aufgaben der einzelnen Abteilungen und
     ihre Zusammenarbeit erklären
1.3 Die Betriebsvorgänge und Arbeitsablä
    ufe im ausbildenden Betrieb beschreiben
```

VI. Results

We considered the best combination for our data set (ABBYY FineReader with GPT-4 omni) for further evaluation. Extensive evaluations would have to be carried out before a generally valid statement could be made about the best combinations. A few of our key findings are the following.

Bad input leads to bad output

One of the findings we had was that a erroneous detection with the standard OCR process will also lead to corrections by UHHGPT that are likely erroneous as well. This becomes clearer when looking at one example. After only using ABBYY FineReader, we got the following detection:

1 id V'V: vJtQ tz 1 rll r G 1 1 kZz \37 \Zz0Gy | l-G-M |ö l ö vG und Arbeitsabläufe im ausbildenden Betrieb beschreiben

(translated into English: 1 id V'V: vJtQ tz 1 rll r G 1 1 kZz \37 \Zz0Gy | l-G-M |ö l ö vG and describe work processes in the training company). The first part could mean anything but based on the semantics, UHHGPT provided the word "Aufbau" (structure) as an appropriate correction. While this word may be fitting, the correct words used where "Wesentliche Geschäftsprozesse" (key business processes) which is semantically not identical to "Aufbau".

A similar problem may also occur when looking at tables. When the regions of the table were not identified correctly, as seen in Section IV-B, and the results are provided as if the table columns were all part of the same sentence, UHHGPT will nevertheless try to find a sentence that may be fitting combining elements of originally different sentences.

The larger the relevant context provided, the better the results

When looking at the results we got after correction with UHHGPT, we noticed that results are generally better when the context provided to it is longer. One example for that was present in the text where a cut-out is shown in Fig. 3. Sometimes, OCR wrongly recognised the chapter number written before every entry, e.g. instead of 1.3 we got 1. When trying to correct that error using UHHGPT, it was able to correct that to 1.3 if the complete table of contents but it did not correct it if only this specific line was given. There are a few factors that may impacts how much context a GPT can take into consideration.

First, GPT has, depending on the specific model, a specific context window. This context window measures how many tokens before the currently generated token are taken into consideration during generation [6]. GPT 4 omni has a context window of 128,000 tokens while GPT 3.5 Turbo only has one of 16,385 tokens [27]. Using the estimates from [28], [38], GPT-3.5 Turbo has a context window of roughly 12,288 words or 24.5 pages while GPT-4 omni has a context window of roughly 96,000 words or around 192 pages. While there are a few documents in our corpus that have more pages than the context window of GPT-3.5 Turbo, GPT-4 omni´s context window should suffice for most.

A second factor to look at is that it may only be useful to provide *relevant* context to the GPT. In one example document about the training of IT specialists, there is a handwritten remark. ABBYY FineReader was unable to detect that correctly and gave "U&ihn s.z". Using UHHGPT, we got "Um UNIX" as a result which may be a correct detection in the context of IT but it was not what was originally written on the page ("vorher 9.2"). Thus, a middle ground between providing enough context and only providing surely relevant context, which may not suffice, must be found.

UHHGPT tends to correct "mistakes" that were no mistakes

In 1998, the German orthography was subject to significant changes. [7] One example is the usage of the letter "ß" (sharp s). Prior to the reform, it was additionally used as a last letter instead of "ss", e.g. today, the spelling "dass" is used but before 1998 it was "daß" (German word for "that"). Other changes included a change from "ph" to "f", e.g. "Photographie" became "Fotografie". There were also a few other changes that are not discussed here.

Our corpus primarily consists of documents written before the reform. Thus, the old spelling was used instead of the new one after 1998. UHHGPT does however correct the old orthography to the new one which does not change the semantics of the text but it may be problematic if one is interested in a one to one digitisation of the original text.

In a similar way, words may be written in all capital letters on purpose, e.g. as a title, or hyphenation was used in a specific way. UHHGPT will "correct" the words to not be written in all capital letters as well as removing all hyphenation. Depending on the specific task, this may be a problem in later steps.

UHHGPT does not always return what it is asked for

When asking UHHGPT to correct the errors using the prompt shown in Section V, it most of the time did what it is asked for. However, in some instances, UHHGPT answered not only with the correction but also an explanation for what is written in the text. Additionally, it sometimes also provided a translation from German into English.

Some quantitative measures

In addition to the qualitative results mentioned above, we got a few quantitative measures. We have taken a few sample pages from the document describing the vocational training for a "Datenverarbeitungskaufmann" from 1995 (see Fig. 3) [3]. Some results are shown in Table I.

TABLE I
SOME RESULTS FOR THE OCR DETECTION AS WELL AS CORRECTIONS BY UHHGPT (GPT-4 OMNI)

Page no.	No. of errors after OCR	No. of errors corrected	No. of errors added
5	5	5	2
7	75	10	28
17	21	5	20
23	6	4	3
28	2	1	3

While GPT-4 omni was able to correct some OCR errors, it also introduced new errors on several pages. For some examples, UHHGPT added more new errors than it has corrected. When using a combination of OCR and LLMs, it should be considered that the results can also contain incorrect corrections when evaluating the data. It is up to the user doing further research with the results to check whether the new errors are relevant or irrelevant for their specific tasks. We analysed some other documents, too, and came to the overall conclusion that the use of GPT-4 omni has made two-thirds of the corrections of the OCR errors.

VII. CONCLUSION AND OUTLOOK

This paper focuses on a corpus of documents pertaining to vocational education research. These regulations were established in the 1920s and have undergone significant evolution over the course of many decades and through various political regimes, including the Third Reich, the GDR, and the FRG. In order to make these documents digitally available for VET research and data science methods, it is crucial to apply OCR methods and digitised (scanned) documents. However, there are different use cases. For the digital archive (Berufearchiv), it is important to have a one-to-one version of the original documents and also provide digital representatives, e.g., in TEI-XML. In the context of data science methods, it is of greater importance to consider the correct language equivalent, given that, for instance, NLP methods do not typically rely on a specific spelling revision but are influenced by poor OCR quality.

In conclusion, for the first use case, it is evident that a human intervention and revision are still necessary. Consequently, future research should focus on collaborative software or further improvements to OCR. In the second use case, our

approaches significantly enhance the quality of the documents, rendering them suitable for NER. However, several critical issues and shortcomings remain. For instance, handwritten artefacts or other alterations to the text are replaced with fictitious texts. With the advancement of OCR, there is a growing need to overcome the challenges presented by historical documents with severe damage. By combining traditional OCR methods with LLMs such as GPT-4 omni, new possibilities for accurately reconstructing and recognising text in damaged documents are emerging. This approach shows promising results in improving text recognition accuracy and preserving the original context of the documents, leading to advances in the preservation and analysis of historical texts.

In our experiments, we have identified the optimal combination for our dataset as ABBYY FineReader with GPT-4 omni (in the form of UHHGPT), which we will further investigate for its effectiveness in the future, in addition, considering other LLMs. As a medium to long-term goal, we can imagine that the approach developed here will be integrated into a RDR (**R**esearch **D**ata **R**epository) in such a way that the combined OCR with LLMs will be offered for the texts in the repository.

REFERENCES

[1] Klassifikation der Berufe, K.: Band 1: Systematischer und alphabetischer Teil mit Erläuterungen (2010)

[2] Bliem, W., Petanovitsch, A., Schmid, K.: Success factors for the Dual VET System. Update (2015)

[3] Bojanowsky, A., Bross, D., Feuerstein, A., Häußler, J., Linde, F., Plattmann, U., Schenk, G., Tumfart, D.: Aktualisierte Ausbildungsplanempfehlungen Datenverarbeitungskaufmann Datenverarbeitungskauffrau. Kuratorium der Deutschen Wirtschaft für Berufsbildung, Adenauerallee 8a, 53113 Bonn (1995)

[4] Bunke, H., Wang, P.S.P. (eds.): Handbook of Character Recognition and Document Image Analysis. World Scientific, Singapore (May 1997). https://doi.org/10.1142/2757

[5] Clausner, C., Antonacopoulos, A., Pletschacher, S.: Efficient and effective OCR engine training. International Journal on Document Analysis and Recognition (IJDAR) **23**, 73–88 (2020)

[6] DeepMind, G.: What is a long context window? Google DeepMind engineers explain (2024), https://blog.google/technology/ai/long-context-window-ai-models/, accessed: 2024-05-18

[7] Dittrich, M.: 25 Jahre Rechtschreibreform: Keiser, Schikoree und Grislibär (2023), https://www.deutschlandfunk.de/rechtschreibreform-deutsche-sprache-100.html, accessed: 2024-05-17

[8] Dobischat, R., Käpplinger, B., Molzberger, G., Münk, D.: Bildung 2.1 für Arbeit 4.0? Springer (2019)

[9] Geis, A.J., Hoffmeyer-Zlotnik, J.H.: Zur Vercodung von Beruf, Branche und Prestige für die DDR, vol. 5. Campus Verl. (1991)

[10] Gessler, M., Howe, F.: From the reality of work to grounded work-based learning in German vocational education and training: Background, concept and tools. International journal for research in vocational education and training **2**(3), 214–238 (2015)

[11] Gewande, W.D.: Historische Entwicklung der staatlich anerkannten Ausbildungsberufe und ihrer Ordnungsmittel von 1934-1999: unter Berücksichtigung der mit deutschen Ausbildungsberufen gleichgestellten österreichischen Lehrberufe und gleichwertigen Facharbeiterberufen aus der ehemaligen DDR. Zentralamt der Bundesanst. für Arbeit, Geschäftsstelle für Veröff. (1999)

[12] Hamann, H.: The German federal courts dataset 1950–2019: From paper archives to linked open data. Journal of empirical legal studies **16**(3), 671–688 (2019)

[13] Harney, K.: Entstehung und Transformation der beruflichen Bildung als Institution–Systemischer Rück-und Ausblick. Bildung und Erziehung **73**(4), 346–357 (2020)

[14] Heliński, M., Kmieciak, M., Parkoła, T.: Report on the comparison of Tesseract and ABBYY FineReader OCR engines. online (2012)

[15] Herkner, V.: Grundzüge der Genese und Entwicklung einer korporatistischen Ordnung von Ausbildungsberufen. Berufsbildung in Wissenschaft und Praxis-BWP **42**(3), 16–19 (2013)

[16] IBM: What Is Optical Character Recognition (OCR)? (2024), https://www.ibm.com/blog/optical-character-recognition/, accessed: 2024-05-17

[17] Koistinen, M., Kettunen, K., Kervinen, J.: How to improve optical character recognition of historical Finnish newspapers using open source Tesseract OCR engine. Proc. of LTC pp. 279–283 (2017)

[18] Kuppe, A.M., Lorig, B., Schwarz, H., Stöhr, A.: Ausbildungsordnungen und wie sie entstehen. Bundesinstitut für Berufsbildung (2015)

[19] Leeuwen, M.v., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven UP (2002)

[20] Li, J., Zhou, H., Huang, S., Cheng, S., Chen, J.: Eliciting the translation ability of large language models via multilingual finetuning with translation instructions (2024), https://arxiv.org/abs/2305.15083

[21] Maier, T.: Die Anwendbarkeit des Erlernten in den wandelnden Bildungs-und Arbeitslandschaften der 1970er-bis 2000er-Jahre. Leverkusen: Verlag Barbara Budrich (2021)

[22] Mani, S.: gImageReader: A Gtk/Qt front-end to tesseract-ocr (2024), https://github.com/manisandro/gImageReader, accessed: 2024-05-18

[23] OCR-D project. https://ocr-d.de/en/, accessed: 2024-05-15

[24] OCR-D Glossary (2024), https://ocr-d.de/en/spec/glossary, accessed: 2024-05-17

[25] OCR4all Workflow (2024), https://www.ocr4all.org/guide/user-guide/workflow, accessed: 2024-05-23

[26] Oliver, D.: Complexity in vocational education and training governance. Research in Comparative and International Education **5**(3), 261–273 (2010)

[27] OpenAI: Models (2024), https://platform.openai.com/docs/models, accessed: 2024-05-18

[28] OpenAI: Tokenizer (2024), https://platform.openai.com/tokenizer, accessed: 2024-05-18

[29] Protsch, P.: Zugang zu Ausbildung: Eine historisch vergleichende Perspektive auf den segmentierten Ausbildungsmarkt in (West-) Deutschland. Tech. rep., WZB Discussion Paper (2011)

[30] Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: OCR4all—An open-source tool providing a (semi-) automatic OCR workflow for historical printings. Applied Sciences **9**(22), 4853 (2019)

[31] Silva, G.P.e., Lins, R.D.: An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. In: 2011 International Conference on Document Analysis and Recognition. pp. 553–557 (2011). https://doi.org/10.1109/ICDAR.2011.117

[32] Universität Hamburg: Research Data Repository. Available: https://www.fdr.uni-hamburg.de/ (2022), accessed: 2024-05-14

[33] VD16: VD 16 digital - Verzeichnis der im deutschen Sprachgebiet erschienenen Drucke des 16. Jahrhunderts der Bayerischen Staatsbibliothek. https://www.digitale-sammlungen.de/de/vd-16-digital-verzeichnis-der-im-deutschen-sprachgebiet/about, accessed: 2024-05-15

[34] VD17: VD 17 - Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts. http://www.vd17.de/, accessed: 2024-05-15

[35] VD18: VD 18 digital - Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 18. Jahrhunderts der Bayerischen Staatsbibliothek. https://vd18.gbv.de/viewer/index/, accessed: 2024-05-15

[36] Wang, Q.F., Yin, F., Liu, C.L.: Improving Handwritten Chinese Text Recognition by Unsupervised Language Model Adaptation. In: 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 110–114 (2012). https://doi.org/10.1109/DAS.2012.46

[37] Wolf, S.: Past meets Present–the history of the German Vocational education and training model as a reflection frame to the prospect of the Egyptian model. Social Dimension and Participation in Vocational Education and Training p. 89 (2017)

[38] WordCounter: Words per Page (2024), https://wordcounter.net/words-per-page, accessed: 2024-05-18