Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

Exploring Historical Labor Markets: Computational Approaches to Job Title Extraction

Raven Adam[1], **Klara Venglarova**[1], **Georg Vogeler**[1]

[1]University of Graz, Graz, Austria

Corresponding author: Raven Adam , raven.adam@uni-graz.at

Abstract

Historical job advertisements provide invaluable insights into the evolution of labor markets and societal dynamics. However, extracting structured information, such as job titles, from these OCRed and unstructured texts presents significant challenges. This study evaluates four distinct computational approaches for job title extraction: a dictionary-based method, a rule-based approach leveraging linguistic patterns, a Named Entity Recognition (NER) model fine-tuned on historical data, and a text generation model designed to rewrite advertisements into structured lists.

Our analysis spans multiple versions of the ANNO dataset, including raw OCR, automatically post-corrected, and human-corrected text, as well as an external dataset of German historical job advertisements. Results demonstrate that the NER approach consistently outperforms other methods, showcasing robustness to OCR errors and variability in text quality. The text generation approach performs well on high-quality data but exhibits greater sensitivity to OCR-induced noise. While the rule-based method is less effective overall, it performs relatively well for ambiguous entities. The dictionary-based approach, though limited in precision, remains stable across datasets.

This study highlights the impact of text quality on extraction performance and underscores the need for adaptable, generalizable methods. Future work should focus on integrating hybrid approaches, expanding annotated datasets, and improving OCR correction techniques to enhance the extraction of structured information from historical texts. These advancements will enable deeper exploration of labor market trends and contribute to the broader field of digital humanities.

Keywords

historical newspapers; job advertisements; job title extraction; NER; occupations dictionary; text-generation; OCR quality effect

I INTRODUCTION

Job advertisements from historical newspapers contain valuable information about history and development of the labor market. However, analyzing text of the advertisements that was turned into machine-readable form through the OCR process (OCRed text [Chiron et al., 2017b]) brings along the challenge of defining their structure and automatically extracting its components, in contrast to digitally-born data. Building on our previous work [Venglarova et al., 2024], we focus on extracting job titles from unstructured historical job advertisements in this paper.

While research concerning modern job advertisements profits from their structure, such as HTML tags, and works with already identified job titles [Colace et al., 2019, Boselli et al., 2017, Zhu et al., 2017], we face the challenge of their definition and automatic extraction. We are comparing several methods for this task.

In comparison with our previous work [Venglarova et al., 2024], we add other models to our evaluation which reach even better performance for the job title extraction task. In addition, we are including an evaluation on an external dataset, to assess the robustness and generalization capacity of the methods used. This is important as we should aim for more general methods and not only corpus-specific ones. While the data in our previous work were manually corrected and cleaned, the real-life applications often include working with non-human-corrected output, whether an automatically post-corrected text or directly a raw OCR output. To account for the text quality influence, we are adding a comparison of our methods' performance on raw OCR, automatically post-corrected and human corrected text.

II RELATED WORK

Modern research on job advertisements often benefits from their structured format and therefore the number of research works that identify job titles is limited. This task was addressed, for instance, by [Bandara et al., 2021] who used rule-based methods to extract job titles from OCRed text, however, with a limited accuracy of 56%. Other approaches include cleaning job titles using a manually created to-delete list [Rahhal et al., 2023] or normalizing job titles manually or semi-automatically [Neculoiu et al., 2016], sometimes while relying on an external taxonomy. Since none of these methods is suitable for our use case, we explore approaches for information extraction.

Named entity recognition (NER) has been widely applied to historical data in previous research [Grover et al., 2008, Ehrmann et al., 2016, Won et al., 2018, Labusch et al., 2019], typically extracting standard entities such as person and location names from historical texts. Although job titles do not belong among standard entities, the results of NER on historical data in general encourages us to develop a custom NER pipeline specifically for extracting job titles.

With the emergence of large language models (LLMs), the NER can also be approached as a text generation or translation task [Keraghel et al., 2024]. To our knowledge, this approach has not yet been tested on historical data; however, it has shown promising results across various domains [Tavan and Najafi, 2022, Wang et al., 2023]. For a more detailed overview of related work in all above-mentioned areas, we point readers to the conference paper [Venglarova et al., 2024].

Building on this foundation, our paper compares several extraction methods, but also introduces their evaluation on an external dataset that was OCRed but not human-corrected, which may impact the performance of our extraction methods. This exploration complements studies like [Rodriguez et al., 2012], who evaluated several NER tools for identifying persons, locations and organizations on two datasets containing raw and manually corrected OCR output. The OCRed text's average word accuracy reached 88.6% and average character accuracy was 93% [Rodriguez et al., 2012]. They conclude that correcting OCRed text did not significantly improve entity recognition or NER tool performance for a certain entity type.

Strien et al. [2020] explored the impact of OCR errors on several tasks common in digital humanities, including NER. Using a corpus of digitized newspapers, they grouped articles into different OCR quality bands. Their findings showed that while NER performance varied across quality bands and entities, OCR errors had a smaller impact on the NER than on tasks such as sentence segmentation or dependency parsing [Strien et al., 2020]. For example, the average F-

	CER	WER	BLEU
Raw OCR	0.0781	0.1972	74.94
Automatically Post-corrected Text	0.0727	0.1589	79.62

Table 1: Quality of raw OCR output and automatically post-corrected text for ANNO corpus. Human-corrected text served as a ground truth.

	Original
Raw OCR	Tu chtiger Ti chler wird aufgenommen. Holzbearbei tungsfabrit Wiiflingseder, 1155 im Innkreis. 5
Automatically Post-corrected Text	Tüchtiger Tischler wird aufgenommen. Holzbearbeitungsfabrik Willingseder, 1155 im Innkreis. 5
Human-Corrected Text	Tüchtiger Tischler wird aufgenommen. Holzbearbeitungsfabrik Wilflingseder, Ried im Innkreis. 135

Table 2: Example of an advertisement text as a raw OCR output, automatically post-corrected and human-corrected text. [Hardworking carpenter will be accepted. Wilflingseder woodworking factory, Ried im Innkreis. 135]. The original image available from Linzen Volksblatt, 8.2.1922, p. 6, https://anno.onb.ac.at/cgi-content/annoshow?call=lvb|19220208|006|100 [cited 12.3.2025]

score for recognizing person entities ranged from 0.87 for the quality band 1 (≥ 0.9 Levenshtein similarity between OCRed text and its ground truth) to 0.63 for the quality bound 4 (≤ 0.7 Levenshtein similarity).

Both studies highlight that recognition results vary across different types of entities. Since job titles are not standard entities, it remains an open question how OCR errors specifically affect their extraction.

III DATASET

Our experiments and evaluation have been conducted on 14 different digitized newspapers from the ANNO corpus [Österreichische Nationalbibliothek, 2021] from 1850-1950 (Fig. 1). From them, we used a subset of job advertisements that we manually annotated, OCRed and manually corrected, and afterwards annotated all job titles within them using the doccano software [Nakayama et al., 2018]. The resulting dataset consisted of 1,486 job advertisements as training data and 637 as testing data. The training and testing dataset are mutually exclusive and were splitted randomly. For more details and examples, we point readers to [Venglarova et al., 2024]. We use the dataset in three versions: as raw OCRed output, as automatically post-corrected text and as human-corrected data. Tab. 1 summarizes their text quality and Tab. 2 shows an example of an advertisement in these three different text-quality settings.

Three different metrics were used to measure text quality. Character error rate (CER) represents the proportion of incorrect characters, due to substitutions, insertions, or deletions, relative to the total number of characters in the correct reference text. A lower CER indicates better performance, with a CER of zero meaning that the recognized text perfectly matches the reference.

Word error rate (WER) measures the proportion of incorrect words in the recognized text compared to the reference text, considering substitutions, insertions, and deletions. Like CER, a lower WER indicates better accuracy, with a WER of zero meaning a perfect match.

Bilingual Evaluation Understudy (BLEU) is a metric used to evaluate the quality of machine-generated text by comparing it to one or more reference texts. It measures how many words or phrases in the generated text match those in the reference, considering both precision and fluency. BLEU assigns higher scores to outputs that closely resemble the reference in wording and structure, with a maximum score of 100 indicating a perfect match.

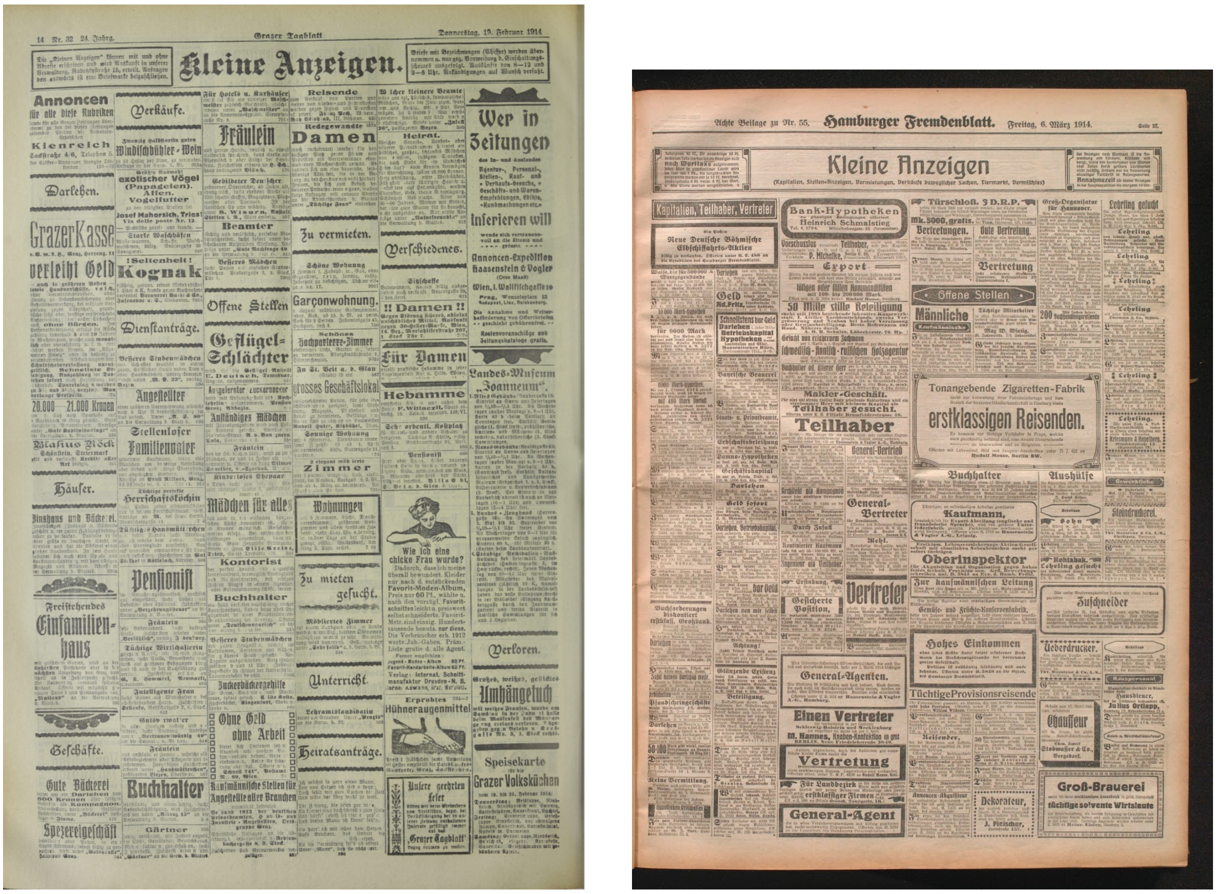
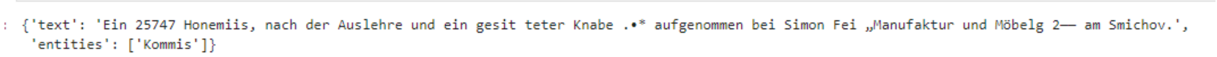


Figure 1: An example page from the ANNO corpus (left) and DDB (right) containing job advertisements. Left image: Grazer Tagblatt, 19.2.1914, p. 14. https://anno.onb.ac.at/cgi-content/annoshow?call=gtb|19140219|14|100.0|0 [cited 11.3.2025]. Right image: Hamburger Fremdenblatt, 6.3.1914, p. 33. https://www.deutsche-digitale-bibliothek.de/newspaper/item/2BZKW5C4ERCLHL7VFCO3BFISZEPJZLFR?issuepage=33 [cited 13.3.2025]

Aiming to assess the impact of text quality and OCR errors on the performance of the different extraction techniques, evaluations were conducted using raw OCR text, automatically corrected text, and human-corrected text. Since the annotated job entities were only available for the human-corrected text, direct matching across the three versions was challenging due to inconsistencies introduced by OCR errors.

To address this, fuzzy matching based on a Levenshtein similarity threshold of 0.6 was employed. This approach allowed each job position in the human-corrected text to be matched to its most likely candidate in the OCR and automatically corrected text. Despite this effort, severe

OCR errors in some cases rendered matching impossible (Fig. 2), leading to a reduction in the dataset size. The final dataset comprised 1882 job advertisements, ensuring that job positions were available across all three text versions for consistent evaluation.



```
: {'text': 'Ein 25747 Honemiis, nach der Auslehre und ein gesit teter Knabe .•* aufgenommen bei Simon Fei „Manufaktur und Möbelg 2— am Smichov.',
   'entities': ['Kommis']}
```

Figure 2: Example of severe OCR errors making the match impossible. The similarity based on Levenshtein distance between *Kommis [clerc]* and *Honemiis* is 0.5 and thus under the similarity threshold

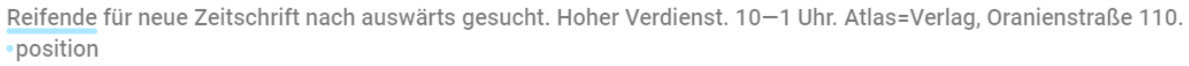
To gain insights about the generalization abilities of the models, we decided to test them also on an external dataset, which was not part of the training nor testing data. For this, we used a subset of a collection of digitized newspapers provided by the German Digital Library (DDB). The entire dataset covers the years 1914 to 1945 but the subset used in this work is from the year 1914 only. As this dataset is fairly new, no ground truth data is currently available and measures for OCR quality such as WER or CER cannot be calculated. After several preprocessing steps to retrieve relevant pages and manual annotation in doccano, the final dataset consisted of 377 job ads. Before the annotation, the raw ocr text was also corrected with a ByT5 model [Löfgren and Dannélls, 2024], fine-tuned on the human-corrected ANNO corpus and the ICDAR2019 dataset.

IV METHODS

4.1 Annotation Process

The first step in annotating job titles within our corpora was to set annotation guidelines. In the ANNO dataset, the job titles were annotated as *positions*, if they explicitly mentioned a position (e.g., *Buchhalter [accountant]*, *Köchin [cook, f.]*) and did not contain any spelling or spacing errors. If errors appeared in the position name, the whole advertisement was excluded from the evaluation. The training dataset includes 1962 annotated positions, while the testing dataset comprises 849 annotated positions.

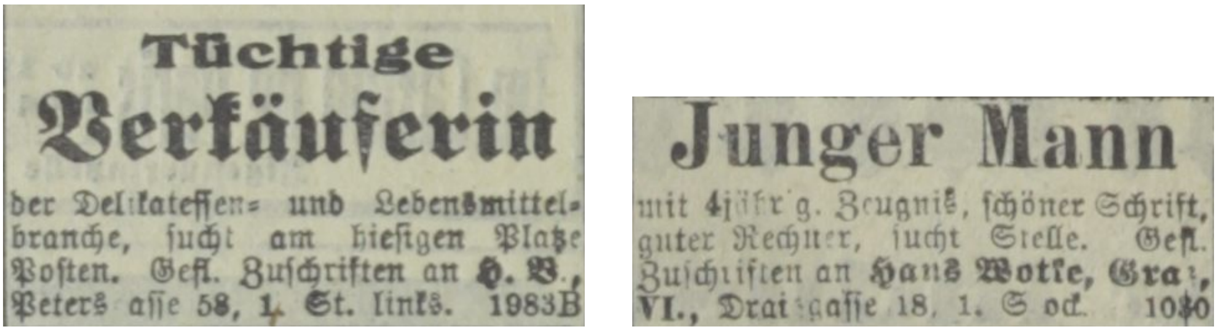
To the DDB dataset, the same rules could not be applied, as the OCR quality was so low that leaving out all the positions containing errors would leave but a very little sample. Also, including positions with errors is more similar to real-life applications, as in reality the large amounts of text can hardly be manually corrected. For this reason, we included all positions that a human annotator could still identify as a position and make a sense of its closest surroundings. That includes also positions with spelling and spacing mistakes (Fig. 3).



Reifende für neue Zeitschrift nach auswärts gesucht. Hoher Verdienst. 10—1 Uhr. Atlas=Verlag, Oranienstraße 110.

Figure 3: Example of an error in annotated positions in the DDB dataset. *Reifende* should be *Reisende* (with a long s). [Travel agent for new newspaper wanted. High earnings. 10-1 o'clock. Atlas=Verlag, Oranienstraße 110.] Doccano interface.

When analyzing preliminary results from the ANNO dataset, we discovered that some models also tended to identify nouns like *woman* as positions in sentences like *A young woman is looking for washing and ironing work*. This led us to reconsider the definition of a job title and introduce a new label *ambiguous* for the DDB dataset to enable us to measure how models identify these nouns while being able to distinguish them from the "standard" positions. This dataset thus comprised 493 *position* entities and 90 *ambiguous* entities (Fig. 4).



Tüchtige

Verkäuferin

der Delikateſſen- und Lebensmittel-
branche, ſucht am hieſigen Platze
Poſten. Geſl. Zuſchriften an H. B.,
Petersgaſſe 58, 1. St. links. 1983B

Junger Mann

mit 4jährig. Zeugnis, ſchöner Schrift,
guter Rechner, ſucht Stelle. Geſl.
Zuſchriften an Hans Wotke, Graz,
VI., Draisgaſſe 18, 1. Stock. 1030

Figure 4: Example of a "standard" (left) and "ambiguous" (right) job title. Left: [Hardworking **saleswoman** in the delicatessen and food branch, looks for a position in the local area. Please write to H. B., Petersgasse 58, 1st floor left. 1983B]. Right: [Young **man** with 4-year reference, nice handwriting, good calculator, is looking for a job. Please write to Hans Wotke, Graz VI, Draisgasse 18, 1st floor. 1030]. Examples from Grazer Tagblatt, 3.18.1914, p. 12, https://anno.onb.ac.at/cgi-content/annoshow?call=gtb|19140318|12|100.0|0 [cited 11.3.2025].

4.2 Post-Correction

Extracting high-quality OCR from historical texts remains a significant challenge, particularly in advertisement sections, which often feature complex layouts and non-standard formatting [Wevers, 2022]. As a result, research has increasingly focused on improving and correcting OCR output using Natural Language Processing (NLP) techniques [Chiron et al., 2017a, Rigaud et al., 2019]. However, the impact of OCR post-correction on downstream tasks, such as NER, remains insufficiently understood, with studies reporting mixed results [Ehrmann et al., 2024].

To assess whether OCR post-correction can enhance job position extraction, we fine-tuned the hmByT5 model on 80% of our human-corrected text data and the DE1, DE2, DE3 and DE7 parts of the ICDAR2019-POCR dataset [Rigaud et al., 2019]. Recent studies indicate that ByT5 variants are particularly well suited for OCR post-correction due to their ability to process text at the character level, which allows for fine-grained corrections [Löfgren and Dannélls, 2024, Debaene et al., 2025, Guan and Greene, 2024].

4.3 Extraction methods

Four distinct techniques were employed to extract job position titles from the advertisements, each leveraging different linguistic and computational methodologies:

- **Dictionary-Based Approach**: A dictionary-based approach was implemented using data from the Historical International Standard Classification of Occupations (HISCO) [Leeuwen et al., 2002]. A dictionary of job titles was constructed by extracting position names and their variants from the HISCO dataset. Job advertisements were then matched against this dictionary, and any exact or approximate matches were identified as job position titles. This approach emphasizes historical relevance and semantic alignment with the HISCO dataset.
- **Rule-Based Approach**: A rule-based method was developed using Part-of-Speech (POS) tagging and linguistic structure analysis. Advertisements were tokenized, and POS tags were assigned to each token using the spaCy 'de_core_news_lg' model [Honnibal and Montani, 2018]. Rules were crafted based on common linguistic patterns observed in job titles. A detailed description of these rules can be found in the previous work [Venglarova et al., 2024] and is also available as code on GitHub (see Code and Data Availability section).

- **Named Entity Recognition (NER) Approach**: A Named Entity Recognition-inspired approach was implemented using the spaCy library to fine-tune 'zeitungs-lm-v1', an Electra model. The model was pre-trained on historical text corpora to capture linguistic nuances specific to the time period. Annotated job advertisements, where job titles were labeled as entities, were used for its fine-tuning. This approach leverages modern transformer-based techniques to detect job titles with high contextual accuracy.
- **Text Generation Approach**: HmByT5, a ByT5 text generation model pre-trained on historical data, was employed to rewrite job advertisements into a structured list of job positions. A curated dataset of historical job advertisements was used to learn the transformation from free-form text to a structured list of positions. The output was then post-processed to ensure comparability with the other approaches.

4.4 Evaluation

Each extraction technique was evaluated using the following metrics:

- **Precision**: The proportion of correctly identified job titles out of all identified titles.
- **Recall**: The proportion of correctly identified job titles out of all actual job titles in the dataset.
- **F1 Score**: The harmonic mean of precision and recall, providing a balanced measure of performance.

Additionally, to assess the generalizability of the techniques, all approaches were also evaluated on the unrelated DDB dataset. This evaluation aimed to determine the robustness of each method when applied to a different linguistic and contextual setting.

V RESULTS

In this section, we present evaluation results for several approaches on different tasks. Table 3 shows the comparison of results of different extraction methods for the raw OCR, automatically post-corrected, and human corrected text from ANNO corpus. We include the evaluation on the training data as an estimation of the difficulty of the task and evaluation on the testing data that the model has not seen during the training process. In the table, the F1-score is presented. The full results containing recall and precision can be found in the Appendix A.

Method	**Raw OCR**		**Automatically Post-corrected Text**		**Human-corrected Text**	
	Train	Test	Train	Test	Train	Test
Dictionary -based	0.517	0.509	0.554	0.552	0.589	0.582
Dictionary -based with Tolerance	0.385	0.379	0.395	0.397	-	-
Rule-based	0.668	0.65	0.681	0.722	0.743	0.714
NER	**0.882**	**0.855**	**0.942**	**0.907**	**0.996**	**0.956**
Text Generation	0.802	0.754	0.829	0.775	0.990	0.922

Table 3: F1-scores for position extraction methods on the texts of different quality.

Table 4 shows the results of job titles extraction on an external dataset that was used to assess our methods' robustness and ability to generalize. The results are presented separately for all annotated entities. In our dataset, 493 'position' entities and 90 'ambiguous' entities are present. In the table, the F1-score is shown. The full results containing recall and precision can be found in the Appendix B. The training and testing datasets are the same for all versions.

Method	All Entities	Position Entities	Ambiguous Entities
Dictionary-based	0.511	0.562	0
Rule-based	0.472	0.466	**0.098**
NER	**0.807**	**0.835**	0.06
Text Generation	0.705	0.766	0.074

Table 4: F1-scores for position extraction methods on the DDB non-human-corrected dataset.

VI DISCUSSION

The performance of all methods remains in consistent order across texts of different qualities, with the lowest performance of the dictionary-based approach and the highest of the custom NER method. All methods perform better with increasing text quality. For the text generation and the NER methods, the results on the training dataset are consistently better than on the testing dataset, which is an expected behaviour.

For the dictionary-based method, a tolerance was introduced through Levenshtein similarity of 0.8. The assumption was that if the lower-quality text contains OCR mistakes, a tolerance would help to perform the matches between the dictionary and job titles within the advertisements. However, the results with the tolerance are worse rather than better. Searching for a cause, we found a number of false-positives introduced by this approach:

***Werkschutz-Wachmänner** sucht grötzeres **Unternehmen**. Bewerbungen mit Lebenslauf, Gehaltsansprüchen und fruühestem Eintrittstermin unter "Werkschutz ND" an den Verlag.*

*[**Plant security guards** sought by a larger **company**. Applications with CV, salary expectations and earliest starting date under 'Werkschutz ND' to the publisher.]*

Annotations: *Werkschutz-Wachmänner [plant security guards]*

Predictions: *Unternehmen [company]*

In this example, the Levenshtein distance allowed for *Unternehmen [company]* to be identified as a job title, because a businessman *[Unternehmer]* is listed in the dictionary.

Despite a decline in performance with lower text quality, the NER approach demonstrated relatively stable results across all three versions of the ANNO dataset. This is in line with previous research showing that transformer based NER is relatively stable across different OCR mistake severities [Strien et al., 2020, Hu et al., 2021]. This consistency of the NER approach can be especially helpful when creating a labeled dataset, as using human-corrected or even automatically corrected text will be easier for annotators than using raw OCR text. In contrast, the performance of the text generation approach notably decreases when used on text with lower quality than it was trained on.

For the DDB dataset, the methods perform in the same order of success as for the ANNO data. Most of the approaches reach lower scores than on the raw OCR ANNO data, which could be however expected due to lower OCR quality of the DDB. Another factor to consider is the possibility of some structural differences between job advertisements in Austrian and German newspapers. In this dataset, both ‘position’ and ‘ambiguous’ entities were annotated. In most cases, the performance is very low for the ambiguous entities, which was expected, as none of the models was specifically trained on them.

While the dictionary-based approach reached the lowest performance, it stays stable across the datasets, as the results for the raw OCR in the ANNO corpus and the DDB dataset are comparable. This welcome behaviour is probably due to the choice of the dictionary, as the HISCO database is not specific to either of our corpora. Potential constraints for further application of this approach to other corpora could be e.g. language variations across time and space.

The rule-based approach had a significant drop of performance for the DDB dataset for the all entities task. However, it has the highest performance for the ambiguous entities from all the approaches. This can be explained by the consistency of the linguistic structure of position and ambiguous entities, even if the rules were built for the position entities only. Extending this result to the NER and text generation approaches, it appears that these methods may place less emphasis on the linguistic structure when developing their representation models. If they did, the identification of ambiguous entities would likely be more consistent. Instead, they appear to focus more on the semantic level, which makes identifying ambiguous entities that also appear in many other contexts much more difficult. This is supported by previous research showcasing that transformer based models excel at semantic representations but show mixed performance when it comes to structural understanding [Pavlick, 2022, Rogers et al., 2020].

VII CONCLUSION

This study highlights the challenges and opportunities in extracting job titles from historical job advertisements using various computational methods. Our findings underscore the importance of considering text quality, linguistic structure, and dataset variability when designing and evaluating extraction techniques.

The NER approach consistently outperformed other methods, demonstrating robustness across text quality variations and datasets. Its ability to learn contextual and semantic representations made it particularly effective and stable, even in the presence of OCR errors. However, its performance on ambiguous entities remains a limitation, suggesting the need for additional training data or model adaptations to better handle such cases. The text generation approach showed promise, particularly on high-quality data, but exhibited greater sensitivity to OCR errors, indicating potential areas for improvement in handling noisy input.

The rule-based approach, while less effective overall, performed relatively better for ambiguous entities than the other approaches, likely due to its reliance on consistent linguistic patterns. This suggests that integrating rule-based heuristics with more advanced models could improve overall performance. The dictionary-based method, despite its limitations, demonstrated stability across datasets, highlighting the utility of domain-specific resources like HISCO for historical data analysis.

Our evaluation on the external DDB dataset revealed the generalization challenges faced by all methods, particularly due to differences in OCR quality and linguistic characteristics. Nonetheless, we were also able to showcase that the trained models generalize reasonably well even to unrelated and low OCR quality data and can help to provide valuable insights into labor market trends and societal changes over time.

VIII ACKNOWLEDGMENTS

We would like to thank the Austrian National Library (ÖNB) for providing the data, Saranya Balasubramanian for a great help with its processing, and Wiltrud Mölzer, Jörn Kleinert, Meike Linnewedel, Clara Hochreiter and Melanie Frauendorfer for their efforts in correcting and annotating it. This work was supported by the FWF under grant number P 35783.

IX CODE AND DATA AVAILABILITY

The code and data containing advertisements text and annotated positions, are available at `https://github.com/JobAds-FWFProject/PositionsExtraction`.

References

RMHD Bandara, HASS Gunasekara, WADS Peiris, WMHC Wijekoon, TS De Silva, SGS Hewawalpita, and HMSC Rathnayake. Information extraction from Sri Lankan job advertisements via rule-based approach. 2021. Publisher: Business Research Unit (BRU).

Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. Using machine learning for labour market intelligence. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10*, pages 330–342. Springer, 2017.

Guillaume Chiron, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. ICDAR2017 Competition on Post-OCR Text Correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 1423–1428, Kyoto, November 2017a. IEEE. ISBN 978-1-5386-3586-5. doi: 10.1109/ICDAR.2017.232. URL `http://ieeexplore.ieee.org/document/8270163/`.

Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4, Toronto, ON, Canada, June 2017b. IEEE. ISBN 978-1-5386-3861-3. doi: 10.1109/JCDL.2017.7991582. URL `http://ieeexplore.ieee.org/document/7991582/`.

Francesco Colace, Massimo De Santo, Marco Lombardi, Fabio Mercorio, Mario Mezzanzanica, and Francesco Pascale. Towards labour market intelligence through topic modelling. 2019.

Florian Debaene, Aaron Maladry, Els Lefever, and Veronique Hoste. Evaluating Transformers for OCR Post-Correction in Early Modern Dutch Theatre. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10367–10374, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-main.690/`.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frederic Kaplan. *Diachronic Evaluation of NER Systems on Old Newspapers*. September 2016.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47, February 2024. ISSN 0360-0300, 1557-7341. doi: 10.1145/3604931. URL `https://dl.acm.org/doi/10.1145/3604931`.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. Named Entity Recognition for Digitised Historical Texts. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, pages 1343–1346, 2008.

Shuhao Guan and Derek Greene. Advancing Post-OCR Correction: A Comparative Study of Synthetic Data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6036–6047, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.361. URL `https://aclanthology.org/2024.findings-acl.361`.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2018.

Yuerong Hu, Ming Jiang, J. Stephen Downie, Glen Layne-Worthey, Ryan C Dubnicek, and Ted Underwood. Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts. In *CHR 2021*, pages 266–279, Amsterdam, 2021.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. A survey on recent advances in named entity recognition, 2024. URL `https://arxiv.org/abs/2401.10825`. _eprint: 2401.10825.

Kai Labusch, Staatsbibliothek Zu, Berlin Kulturbesitz, Clemens Neudecker, and David Zellhöfer. BERT for Named Entity Recognition in Contemporary and Historical German. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, Erlangen, Germany, October 2019.

Marco H. D. van Leeuwen, Sören Edvisson, Ineke Maas, and Andrew Miles. *HISCO : Historical International Standard Classification of Occupations*. 2002. ISBN 90-5867-196-8. URL https://iisg.amsterdam/en/data/data-websites/history-of-work.

Viktoria Löfgren and Dana Dannélls. Post-OCR Correction of Digitized Swedish Newspapers with ByT5. In Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz, editors, *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 237–242, St. Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.latechclfl-1.23/.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text Annotation Tool for Human, 2018. URL https://github.com/doccano/doccano.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. *Learning Text Similarity with Siamese Recurrent Networks*. January 2016. doi: 10.18653/v1/W16-1617.

Ellie Pavlick. Semantic Structure in Deep Learning. *Annual Review of Linguistics*, 8(1):447–471, January 2022. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-031120-122924. URL https://www.annualreviews.org/doi/10.1146/annurev-linguistics-031120-122924.

I. Rahhal, K. M. Carley, I. Kassou, and M. Ghogho. Two Stage Job Title Identification System for Online Job Advertisements. *IEEE Access*, 11:19073–19092, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3247866.

Christophe Rigaud, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. ICDAR 2019 Competition on Post-OCR Text Correction, September 2019. URL https://zenodo.org/record/3459116. Conference Name: 15th International Conference on Document Analysis and Recognition (ICDAR) Publisher: Zenodo.

Kepa J. Rodriguez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. *Comparison of Named Entity Recognition tools for raw OCR text*. September 2012. doi: 10.13140/2.1.2850.3045.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, December 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00349. URL https://direct.mit.edu/tacl/article/96482.

Daniel Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara Mcgillivray, and Giovanni Colavizza. *Assessing the Impact of OCR Quality on Downstream NLP Tasks*. February 2020. doi: 10.5220/0009169004840496.

Ehsan Tavan and Maryam Najafi. MarSan at SemEval-2022 Task 11: Multilingual complex named entity recognition using T5 and transformer encoder. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1639–1647, Seattle, United States, June 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.226. URL https://aclanthology.org/2022.semeval-1.226.

Klara Venglarova, Raven Adam, and Georg Vogeler. Extracting position titles from unstructured historical job advertisements. In Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni, editors, *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 75–84, Miami, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.nlp4dh-1.8.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. GPT-NER: Named Entity Recognition via Large Language Models, 2023. URL https://arxiv.org/abs/2304.10428. _eprint: 2304.10428.

Melvin Wevers. Mining Historical Advertisements in Digitised Newspapers. In *Digitised Newspapers – A New Eldorado for Historians?*, pages 227–252. De Gruyter Oldenbourg, December 2022. ISBN 978-3-11-072921-4. doi: 10.1515/9783110729214-011. URL https://www.degruyter.com/document/doi/10.1515/9783110729214-011/html.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5, 2018. ISSN 2297-2668. doi: 10.3389/fdigh.2018.00002. URL https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2018.00002.

Yun Zhu, Faizan Javed, and Ozgur Ozturk. Document embedding strategies for job title classification. In *The Thirtieth International Flairs Conference*, 2017.

Österreichische Nationalbibliothek. ANNO Historische Zeitungen und Zeitschriften, 2021. URL https://anno.onb.ac.at/.

A APPENDIX RESULTS FOR ANNO DATASET

	Raw OCR					
Method	**F1 Score**		**Recall**		**Precision**	
	Train	Test	Train	Test	Train	Test
Dictionary -based	0.517	0.509	0.521	0.538	0.513	0.484
Dictionary -based with Tolerance	0.385	0.379	0.367	0.372	0.406	0.387
Rule-based	0.668	0.65	0.674	0.653	0.661	0.647
NER	**0.882**	**0.855**	**0.815**	**0.787**	**0.961**	**0.936**
Text Generation	0.802	0.754	0.793	0.739	0.810	0.769

Table 5: F1-score, recall and precision for different methods for the task of job titles extraction on the raw OCR text from the ANNO dataset.

	Automatically Post-corrected Text					
Method	**F1 Score**		**Recall**		**Precision**	
	Train	Test	Train	Test	Train	Test
Dictionary -based	0.554	0.552	0.606	0.629	0.51	0.492
Dictionary -based with Tolerance	0.395	0.397	0.393	0.408	0.398	0.387
Rule-based	0.681	0.722	0.612	0.731	0.769	0.713
NER	**0.942**	**0.907**	**0.911**	**0.866**	**0.975**	**0.953**
Text Generation	0.829	0.775	0.829	0.769	0.829	0.781

Table 6: F1-score, recall and precision for different methods for the task of job titles extraction on the automatically post-corrected text from the ANNO dataset.

	Human-corrected Text					
Method	**F1 Score**		**Recall**		**Precision**	
	Train	Test	Train	Test	Train	Test
Dictionary -based	0.589	0.582	0.454	0.446	0.837	0.838
Dictionary -based with Tolerance	-	-	-	-	-	-
Rule-based	0.743	0.714	0.754	0.721	0.733	0.708
NER	**0.996**	**0.956**	**0.995**	**0.939**	**0.997**	**0.973**
Text Generation		0.922		0.913		0.932

Table 7: F1-score, recall and precision for different methods for the task of job titles extraction on the human-corrected text from the ANNO dataset.

B APPENDIX RESULTS FOR DDB DATASET

Method	**F1 Score**	**Recall**	**Precision**
Dictionary-based	0.511	0.424	0.641
Rule-based	0.472	0.393	0.589
NER	**0.807**	**0.743**	**0.882**
Text Generation	0.705	0.609	0.838

Table 8: F1-score, recall and precision for different methods for the task of job titles extraction from the DDB dataset. Results for both labels ‘position’ and ‘ambiguous’.

Method	**F1 Score**	**Recall**	**Precision**
Dictionary-based	0.562	0.5	0.641
Rule-based	0.466	0.416	0.529
NER	**0.835**	**0.783**	**0.893**
Text Generation	0.766	0.704	0.841

Table 9: F1-score, recall and precision for different methods for the task of job titles extraction from the DBB dataset. Results for the label ‘position’.

Method	**F1 Score**	**Recall**	**Precision**
Dictionary-based	0	0	0
Rule-based	**0.098**	**0.264**	0.06
NER	0.06	0.145	0.038
Text Generation	0.074	0.092	**0.121**

Table 10: F1-score, recall and precision for different methods for the task of job titles extraction from the DDB dataset. Results for the label ‘ambiguous’.