Extracting position titles from unstructured historical job advertisements

Klara Venglarova
University of Graz
klara.venglarova@uni-graz.at

Raven Adam
University of Graz
raven.adam@uni-graz.at

Georg Vogeler
University of Graz
georg.vogeler@uni-graz.at

Abstract

This paper explores the automated extraction of job titles from unstructured historical job advertisements, using a corpus of digitized German-language newspapers from 1850-1950. The study addresses the challenges of working with unstructured, OCR-processed historical data, contrasting with contemporary approaches that often use structured, digitally-born datasets. We compare four extraction methods: a dictionary-based approach, a rule-based approach, a named entity recognition (NER) mode, and a text-generation method. The NER approach, trained on manually annotated data, achieved the highest F1 score (0.944 using transformers model trained on GPU, 0.884 model trained on CPU), demonstrating its flexibility and ability to correctly identify job titles. The text-generation approach performs similarly (0.920). However, the rule-based (0.69) and dictionary-based (0.632) methods reach relatively high F1 Scores as well, while offering the advantage of not requiring extensive labeling of training data. The results highlight the complexities of extracting meaningful job titles from historical texts, with implications for further research into labor market trends and occupational history.

1 Introduction

Historical job advertisements provide unique information about the history and development of the labor market. Analyzing the positions offered and sought over time offers insights into temporal and regional differences and development, as well as into social aspects, such as gender-specific job offers. The first step in such an analysis is the extraction of job titles. However, when using historical data from digitized newspapers, rather than digitally-born structured data, the automatic extraction of job titles proves to be a non-trivial task.

In the JobAds Project (FWF P35783), we study historical job advertisements from digitized newspapers from the ANNO corpus (Österreichische Nationalbibliothek, 2021). The advertisements are predominantly in German, and our defined time span is 1850-1950. The newspaper pages were initially obtained in the form of images and transformed into textual data by conducting the processes of page segmentation, optical character recognition (OCR), and automatic post-correction based on manually transcribed ground truth. Afterwards, the job advertisements were extracted. A result of such a pipeline is a corpus containing tens of thousands machine-readable yet unstructured job advertisements.

Contemporary research often works with modern, digitally-born data, and usually benefits from their structure, such as HTML tags, to identify position titles. Modern research in the context of job advertisements addresses challenges such as extracting or grouping requirements in the job ads, e.g. (Gnehm et al., 2022; Ternikov, 2022; Grüger & Dr. Schneider, 2019; Wowczko, 2015; Litecky et al., 2010), automated matching process between the position requirements and the skills of a candidate written in their CV, e.g. (Fernández-Reyes & Shinde, 2019; Sayfullina et al., 2018; S. Chala et al., 2017; Guo et al., 2016; J. Malinowski et al., 2006), job advertisements classification/categorization, e.g. (Gnehm & Clematide, 2020; Boselli et al., 2018; Malherbe et al., 2015; Amato et al., 2015), and job title classification, e.g. (Colace et al., 2019; Boselli et al., 2017; Zhu et al., 2017); with some works covering more than one research focus. In contrast, we face the challenge of working with unmarked text, and need to find an automated way to extract this information.

In this paper, we present a comparison of four approaches to identification of position titles in historical job advertisements. On the one hand, we consider dictionary-based and rule-based approaches, which do not require a time-consuming creation of

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 75–84

an annotated training dataset. On the other hand, we use a machine-learning (ML) approach, specifically named entity recognition (NER) and text-generation models training. This requires manually annotated training data, but yields better results and can recognize position titles beyond those seen in the training data. Please note that in this paper, we aim to extract position titles appearing inside the advertisement, in contrast to some of the aforementioned research works which aimed to classify jobs into the occupation categories.

The following section summarizes existing research related to job title identification and information retrieval from unstructured text. Section 3 describes our dataset and the process of job titles annotation. Section 4 presents in detail various approaches we used to identify the job titles and the evaluation methodology, while section 5 presents and discusses the results. Section 6 concludes this paper.

2 Related Work

While most of the modern research benefits from the structure of ads to extract job titles, as exemplified above, some research addresses the challenge of their extraction nevertheless. One reason for this can be the noisy information included within the job title tags, such as the name of the company, or the need to work with ads from newspapers that were obtained in the form of the image.

Rahhal et al. (2023) develop a methodology for matching a job ad with its standardized occupation in French language. Although the job title is included within a dedicated HTML tag, the authors further process it because other words, such as 'looking for', are sometimes included. They remove these extraneous words based on a manually created to-delete list to obtain cleaned job titles.

Bandara et al. (2021) work with unstructured text, as they scrape job ads in the form of images from job web portals and newspapers. They apply an OCR process to convert them into machine-readable text. Aiming to create a structured dataset, they extract information such as the position name, skills, company name, and contact information. They use a rule-based approach, matching regex patterns or phrases. However, the accuracy of job title identification is only 56% (Bandara et al., 2021, p. 148). The authors do not explicitly state which specific regex patterns were used for job titles extraction.

Neculoiu et al. (2016) focus on job title normalization, where job titles are normalized according to a predefined set of occupations. Researchers dealing with this question also need to extract job titles; however, they often rely on external taxonomies, structured text, or manual labeling. In (Neculoiu et al., 2016, p. 152), "the job titles were manually and semiautomatically collected from resumes and vacancy postings."

Not having found a suitable approach in the existing literature, we focus on other ways to extract job titles from the advertisements. Information extraction (IE) is the process of automatically extracting entities, objects, and their roles from text, often within a specific domain or topic (Hobbs & Riloff, 2010), with named entity recognition (NER) being one of the common technologies used (Tjong Kim Sang & De Meulder, 2003; Collins & Singer, 2002; Cucerzan & Yarowsky, 2001). While the standard entities in NER models are typically proper names, locations, or dates, custom NER models can be trained to recognize new entities. Therefore, we can consider job titles entities, and having created appropriate training data, use the same approach to extract job titles from the unstructured text.

The emergence of large language models (LLMs) has also opened up the possibility to approach NER as a text generation or translation task (Keraghel et al., 2024). While this approach has, to the best of our knowledge, not been evaluated on historical data, it has seen promising results in a variety of fields, e.g., (Tavan & Najafi, 2022; Wang et al., 2023), and we shall include this method in our evaluation.

While no work specifically addresses position extraction from historical data, several works deal with historical data and NER. Grover et al. (2008) addressed recognition of person and place names within a digitized corpus of British parliamentary proceedings from 1685–1691 and 1814–1817 using a rule-based approach. Working with the output of an OCR software, they reach total f-scores from 70.35 to 76.94 on individual datasets (Grover et al., 2008, p. 1346). The main challenge were the OCR mistakes, namely the noise, misrecognition of characters and issues with separating the text from the marginal notes.

Won et al. (2018) focus on toponyms in two collections of historical letters, one collection in early-modern English, another in modern English. As obstacles they mention e.g. "language changes over time, spelling variations, OCR errors, sources

written in multiple languages, and general ambiguity in language use" (Won et al., 2018, p. 2). The authors do not train new models but evaluate the performance of existing modern tools for the task of location extraction, including two different types of pre-processing and experimenting with re-writing early-modern English text into modern English. They evaluate five different NER systems and an ensemble method which works based on a voting system among the individual NER models. All experiments reached a best minimum F1 score of about 70 (Won et al., 2018, p. 8).

Labusch et al. (2019) trained a pre-trained BERT model for a NER task in contemporary and historical German corpora, containing entities of location, organisation and person. Working with historical corpora is hindered through the less standardized languages and errors in the OCR. Combining unsupervised pre-training on historical German corpus with supervised pre-training using contemporary German NER ground-truth, they achieve the highest F1 score of 84.6 on historical data for 5-fold cross validation.

3 Dataset

Using digitized historical newspapers from the ANNO corpus, we manually annotated, OCRed and corrected several thousands job advertisements spanning the period 1850-1950 from 14 different newspapers. Within a subset of these job ads, we annotated position names using the doccano software (Nakayama et al., 2018), yielding 1,486 job advertisements as training data and 637 as testing data. The split into training and testing dataset was random.

Ads that did not explicitly mention a position or contained errors, such as inconsistent spacing, were mostly not included within the datasets for the evaluation task (Tab. 1), however, after the training and evaluation during a human control we discovered a small number of ads containing errors that were included in the dataset by mistake. Additionally, we standardized the text by replacing the *long s* with the letter *s* in all ads. The training data was used to train the NER and text-generation models and also served as a basis for creating part-of-speech (POS) and syntax rules, as well as for adding entries to the dictionary, both established through human observations. The testing data was consistently used for evaluation purposes.

Although all duplicate ads were removed and the training and testing datasets are mutually exclusive, we cannot exclude the possibility that very similar ads appear in both datasets, given the homogeneity of ads and their structure defined by their genre. Each advertisement could contain more than one position offered/sought for.

4 Methods

Based on preliminary results, four approaches were selected and compared: a dictionary-based approach, a rule-based approach, a NER and a text generation approach.

4.1 Dictionary-based Approach

The dictionary-based approach searches for position titles based on the database of historical occupations HISCO[1] (Leeuwen et al., 2002). The HISCO (historical international classification of occupations) database is based on the coding of 1,000 most frequent male and female occupational titles in datasets from Belgium, Britain, Canada, France, Germany, the Netherlands, Norway and Sweden. It contains titles included in parish and civil registration documents (International Institute of Social History, 2023). The dictionary of German occupations contains 1297 job titles and was subsequently further enriched based on our observations by collocations (e.g. *Mädchen für alles* [*girl servant*, lit.*girl for everything*], *Stütze der Hausfrau* [*housewife's help*]) and positions related to apprenticeship (e.g. *Lehrmädchen* [*apprentice* f.], *Praktikant* [*intern*]), which we also aim to identify. Every exact match of the collocations was identified as a position.

To overcome difficulties of different spelling variants (*Commis* and *Kommis* [*assistance*]), we lemmatized every entry in the dictionary as well as each token in the advertisement using the DTA::CAB web-service (Jurish, 2012)[2] which serves for an 'error-tolerant linguistic analysis for historical German text'. If a lemmatized token matched a position from the dictionary, regardless of the upper/lowercase, the entire word containing this string was identified as a position. Even when this approach introduces some false positives (e.g., identifying *Architektur* [*architecture*] because it contains *Architekt* [*architect*]), it helps to identify true positives that are not in the dictionary (e.g., *Steinbrechermeister* [*stone crusher mas-*

[1] https://iisg.amsterdam/en/data/data-websites/history-of-work [15.6.2024]

[2] https://www.deutschestextarchiv.de/demo/cab/ [26.8.2024]

Advertisement text	Part of the corpus	Annotated entities
Kinderliebendes Mädchen zu 2 Kindern und Mithilfe dringend gesucht. Zuschriften unter „Kinderliebend 2148“ an Rasteiger, Joanneumring 6. 1559 [Child-loving girl urgently wanted to 2 children and assistance. Send letters under ‘Kinderliebend 2148’ to Rasteiger, Joanneumring 6. 1559]	No (no position title)	-
Suche eine tüchtige **Wirt schafterin**. 35—40 J. spätere Ehe nicht ausgeschlossen. Un ter „Ehrlich 2270“ Neue Zeit [Looking for a hard-working **housekeeper**. 35-40 yrs. later marriage not excluded. Under Sincere 2270’ Neue Zeit]	No (wrong spacing)	-
Ein **Commis** in einer Eienhandlung wird acceptirt. 2936 [An **assistance** will be accepted in an ironmonger. 2936]	Yes	Commis
Maschinenschlosser resp. **Automatenschlosser** wird gesucht. Vorzustellen Ing. Gasser-Steiner Graz Strauchergasse 16. [**Machine locksmith** or **automatic locksmith** is wanted. To be introduced at Ing. Gasser-Steiner Graz Strauchergasse 16.]	Yes	Maschinenschlosser, Automatenschlosser

Table 1: Example of annotated position names in job advertisement and of advertisements excluded from our dataset.

ter], *Tapezierergehilfen* [*paperhanger's assistant*], *Weißnäherin* [*seamstress*]), often because they are composed of more words that specify the position. The number of positions identified is equal to the number of matches.

4.2 Rule-based Approach

The rule-based approach benefits from the strong linguistic structure of job ads. First, the same collocations are searched for as in the dictionary-based approach (e.g., *Mädchen für alles* [*girl servant*, lit. *girl for everything*], *Stütze der Hausfrau* [*housewife's help*]). If a match is found, the collocations are identified as positions. If no match is found, the word ‘ *als* ’ [*as*] (or ‘*Als* ‘ [*As*]) is searched with whitespaces as stated. If this word is found, the first noun after it is identified as a position, e.g.:

*Suche Stelle **als** Hausgehilfin in Bäckerei od. Gastwirtschaft.* [*Looking for a job **as** a maid in a bakery or restaurant.*]

If this search is unsuccessful, the search continues for the word ‘*stelle*’ [*position*]. Note that in German, *stelle* beginning with lowercase can only be found in the middle of a word. If this is found, the string containing it is considered a position, e.g.:

*Alleinstehende Frau mit kl. einjähr. Buberl bittet um Hausmeister**stelle**.* [*Single woman with a small one-year-old boy asks for a caretaker **position**.*]

If none of these conditions is met, the ad is searched for the first noun which is subject, root, or conjunct in the sentence. These syntactic roles were chosen based on observations in training data, including the observed miss-classifications caused probably by concise and archaic language. This noun is predicted as a position. If it is followed by ‘*und*’ [*and*] or ’*oder*’ [*or*] and the word after this conjunction is a noun, this second noun is considered another position.

The advantage of this approach is its independence from an external dictionary, spelling variations and words that do not appear in the dictionary. The main drawback is handling advertisements that offer more than one position, as it is hard to define rules that include multiple true positives without introducing false positives. In the testing data, 514 instances contain one single position, while 123 of them contain more than one position. For the POS tagging and dependency parsing, the SpaCy library (Honnibal & Montani, 2017) with the ‘de_core_news_lg’ model was used.

4.3 Named Entity Recognition Approach

The NER approach treats the positions in the advertisements as named entities and attempts to identify them accordingly. As positions are not among

the standard entities, a custom model needs to be trained.

We trained two different models, one with the Token2Vec architecture, and one with a transformer-based architecture. The spacy.Tok2Vec.v2 model contains two steps: it creates context-independent word vector representation, and it encodes context into the embeddings, using architecture like a CNN, BiLSTM or transformer (SpaCy, n.d.). In comparison, the spacy-transformers.TransformerModel.v3 uses transformer models from the HuggingFace transformers library to create more advanced, context-aware embeddings, leading to better performance in tasks like NER (SpaCy, n.d.). Training on GPU allows for faster training times, particularly for transformer-based models, which involve complex matrix operations (Kelleher, 2019, pp. 92–94), and they reach generally higher performance than training on CPU.

For the training, we used default SpaCy settings. In both cases, the language is set to 'German', and the optimization is set to efficiency. In the first case, we set the hardware to CPU, while in the second case, we select GPU. For further details on the training parameters, we point readers to SpaCy official documentation[3].

In both cases, the NER model was trained on the training dataset and evaluated on the testing dataset, as specified in the Dataset section. The advantage of this approach is some ability to generalize; however, the disadvantage is the need for manual training data creation, which can be time-consuming, and may limit model's effectiveness if the training data is not selected appropriately.

4.4 Text Generation Approach

Whereas the NER approach identifies and extracts entities within text, the text generation approach creates new text that directly facilitates the identification of relevant entities. In the chosen approach a given text, e.g. "Machine locksmith or automatic locksmith is wanted." is rewritten into "<Machine locksmith> <automatic locksmith>". If no position is found within a given text, it is rewritten into "none".

We used a hmByT5 model[4] as our base model due to two major reasons. The first one being that the hmByT5 models are all trained on multilingual historical data, which overlaps with the time period of our dataset. The second one being that ByT5 encoded text byte-wise instead of the word or subword level. Therefore, the model requires no internal vocabulary and is more adaptable to words that were not included in the base model training.

Because a significant number of job ads exceeds the maximum encoding length of the chosen model, all ads concerned were split into segments of up to 120 bytes for training and evaluation. Afterwards the model was trained for 15 epochs and the epoch with the highest sacreBLEU score was chosen as the final model. Due to splitting up ads into smaller segments, the model is evaluated in two different ways (Tab. 2). First, based on the prediction for each segment. Since not all segments contain positions, properly predicting "none" as a result is included when calculating evaluation metrics. The second evaluation re-merges the predictions from all segments into the original ads and evaluates only the predicted positions without including "none" predictions, as this is most comparable to the NER approach.

4.5 Evaluation

The evaluation involves a pre-processing step, in which we aim to standardize both the annotated positions, and the predicted ones, to avoid penalizing a model for e.g., including white spaces. First, we delete the word 'stelle' [*position*] if it is included, e.g. *Hausmeisterstelle* [*caretaker position*] becomes *Hausmeister* [*caretaker*]. Afterwards, both annotated and predicted positions are lemmatized using the DTA::CAB web-service (for details, see section 4.1).

In the next step, the two lists of annotated and predicted positions are compared for each advertisement. Certain tolerance is included by allowing Levenshtein distance (Levenshtein, 1965) of 0.1 between the two strings in order not to penalize the generative approach if it makes minor mistakes, such as generating (position> instead of <position>. The number of True Positives, False Positives and False Negatives is identified as follows:

- **True Positives (TP)**: The model correctly predicts a position that is present in the list of annotated positions.
- **False Positives (FP)**: The model predicts a position that is not present in the list of annotated positions.
- **False Negatives (FN)**: The model misses a

[3] https://spacy.io/usage/training [29.7.2024]

[4] https://huggingface.co/hmbyt5-preliminary/byt5-small-historic-multilingual-span20-flax [25.8.2024]

Advertisement segment	Model output	Aggregation of entire output
Aelterer Herr, **Beamter** i. P. gesund und rüstig, alleinstehend sucht Posten in Schloß, Kloster [Older gentleman, **civil servant**, healthy and sprightly, single seeks position in castle, monastery]	<Beamter>	<Beamter>, <Ausseher>, <Pförtner>, <Hausgehilfe>
als **Ausseher**, **Pförtner**, **Hausgehilfe** zu Gartenarbeiten auch auswärts. Gute Zuschrifen [as **external**, **doorman**, **housekeeper** for gardening work also away from home. Good applications]	<Ausseher>, <Pförtner>, <Hausgehilfe>	<Beamter>, <Ausseher>, <Pförtner>, <Hausgehilfe>
unter: "Vertrauenswürdig und verlässlich Nr 85368" an das Welt-Blatt. Wien, I. Schulerstraße. [under: "Trustworthy and reliable no. 85368W to the Welt-Blatt. Vienna, I. Schulerstrasse.]	none	<Beamter>, <Ausseher>, <Pförtner>, <Hausgehilfe>

Table 2: Example of segmented job advertisement along with text generation model output and aggregation of position names found for the whole add without including "none" predictions. Evaluation is performed directly on the outputs as well as the aggregated predictions.

position that is present in the list of annotated positions.

Using this information, we calculate F1 Score, Recall, and Precision (Powers, 2011), where the metrics are calculated as follows:

- **F1 Score**: 2 * (Precision * Recall) / (Precision + Recall)
- **Recall**: TP / (TP + FN)
- **Precision**: TP / (TP + FP)

This evaluation is identical for all four approaches, making their scores comparable.

5 Results and Discussion

Table 3 presents the results using F1 Score, Recall, and Precision for the above described methods on the testing dataset.

The NER approach (Fig. 1) reached the highest F1 score of 0.944 using the transformer architecture, resp. 0.884 when trained on CPU on the testing dataset, indicating that this method is more flexible than the rule-based and dictionary-based. However, a significant drawback of this approach is its reliance on a manually created training set.

The generative approach shows comparable performance with an F1 score of 0.920 when evaluated based on segments and 0.902 when evaluated on the aggregated results for entire ads. Since the only difference between the two evaluation approaches is the exclusion of "none" predictions when aggregating, the decrease from the segment based score

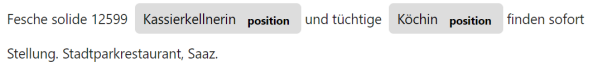

Figure 1: Example of positions identified by the NER model. [*Good-looking, solid 12599 cashier waitress and hard-working cook are wanted immediately. City park restaurant, Saaz.*]

to the ad based score indicates that the model is slightly better suited to determine that no positions are mentioned in a line of text instead of extracting all mentioned positions. A preliminary check of prediction mistakes has, however, shown that some false positives occur from identifying ambiguous words such as *Mann* [*man*], *Mädchen* [*girl*] or *Französinnen* [*French women*], which are also used to advertise positions but were not labeled in our dataset as well as words with misplaced spacing, such as *Wirt schafterin* [*ho usekeeper*], which was also not included as a labeled position. This behavior indicates potentially interesting and useful generalization ability.

The rule-based approach appears to be well-suited for this task, given the strong linguistic structure of job advertisements. It is also immune to certain spelling variations or typographic mistakes, as long as they do not include spacing errors, or do not hinder the correct POS classification. The problematic aspect of this approach is when more than one position is advertised within an ad, leading to either incomplete detection of all positions or the introduction of false positives in other ads. Another technical obstacle is the misclassification

Method	F1 score	Recall	Precision
Dictionary-based approach	0.632	0.646	0.617
Rule-based approach	0.690	0.613	0.789
NER approach (CPU)	0.884	0.866	0.903
NER approach (GPU)	0.944	0.932	0.956
Text Generation approach (segment)	0.920	0.918	0.922
Text Generation approach (whole ad)	0.902	0.894	0.909

Table 3: Results for different approaches for predicting position names on the testing dataset.

of parts-of-speech or dependencies due to archaic and elliptic language used within historical job advertisements, while the modern tools are generally designed for modern data and contemporary language. They may also not handle advertisements containing a large number of abbreviations, which is often the case because every line in a newspaper was costly.

The dictionary-based approach reached the lowest F1 score. While lemmatizing/standardizing of the tokens mainly solves the issue of spelling variants, it is time-consuming to lemmatize the entire text of the advertisement. However, this approach without lemmatization reaches lower score as usually only one form is present in the dictionary, while the text contains several variants (e.g., *Commis* and *Kommis* [*assistance*], *Kontoristin* and *Comptoiristin* [*clerk* f.]). Additionally, some professions are simply not present in the dictionary, especially those that were less common or highly specialized. This approach also fails when typographic errors are present in the text (*Kö chin* with line-breaker in the original text instead of *Köchin* [*cook* f.]). Moreover, all mentioned positions are identified, even if they appear as part of a name of a street, or if they are not the primary focus of the advertisement but just mentioned within it, e.g.:

Kinderarzthilfe. mit zweijähriger Praxis, sucht Stelle bei ***Arzt*** *oder in einem Laboratorium.* [*Pediatric assistant with two years of experience is looking for a position with a* ***doctor*** *or in a laboratory.*]

Our dataset contained advertisements in their ideal form, i.e., manually corrected with only occasional typographic errors. In reality, the thousands of ads can only be post-corrected automatically rather than manually, and more errors will thus be present in the data. This will mostly affect the dictionary-based approach, which needs exact matches with correct spelling, and partially the rule-based approach, which is in theory immune to the spelling variations but fails to correctly identify the POS and syntax dependencies if too many errors occur.

The last consideration is the ambiguous nature of what a job title in historical job advertisements is. Many ads contain words like *Mann* [*man*], *Mädchen* [*girl*] without specifying a clear job title, although it is implicitly understood that e.g., the girl is wanted to help in the household. Another example is *Französinnen* [*French women*], mentioned alongside other job titles for women, where it is implied that they are sought to teach the French language or to provide companionship while conversing in French (Fig. 2). These advertisements were not included in our dataset for this evaluation task, however, they appear commonly in the corpus.


Tüchtige
Erzieherinnen,
Stundenlehrerinnen, Kindergärtnerinnen, Französinnen, Gesellschafterinnen, Wirthschafterinnen und Bonnen
empfiehlt und placirt bestens
Regina Kohn, Placirungsbur.,
Herrengasse 4, 1. Stock.

Figure 2: [*Regina Kohn, Placement Office, Herrengasse 4, 1st floor, recommends and places competent educators, class teachers, kindergarten teachers,* ***French women****, companions, housekeepers, and maids.*] Source: Prager Tagblatt, 23.8.1896, p. 30, https://anno.onb.ac.at/cgi-content/anno?aid=ptb&datum=18960823&seite=30

Further research could help to address these limitations by focusing on the semantic context of the advertisements, perhaps using techniques like average embeddings to represent the average meaning

of the whole advertisement. This approach could help overcome the challenges posed by the ambiguous and context-dependent nature of job titles in historical advertisements. Also, the potential of text-generation methods shall be further investigated.

6 Conclusion

Historical job advertisements offer many opportunities to study the transformation of the labor market. The first step in analyzing these ads is the extraction of job titles, which allows for the exploration of positions offered and sought, their frequency analysis, and temporal and regional variations. In this study, we compared four approaches for job title extraction: a dictionary-based approach, a rule-based approach, a NER approach, and a text-generation approach. The NER approach achieved the highest F1 score of 0.944 when a transformer-based model was trained on GPU, resp. 0.884 when a model was trained on CPU, as we treated position names as entities and trained a model specifically for their identification. The generative model follows closely with a F1 score of 0.920 and offers potential viability for identifying ambiguous positions. The rule-based (0.69) and dictionary-based (0.632) approaches had lower F1 Scores but offered the advantage of not requiring the time-consuming creation of a training dataset.

While extracting position names is a crucial first step, two important considerations remain:

1. Not all job advertisements explicitly state the position being offered or sought.

2. As noted in (Wowczko, 2015, p. 36), "job titles can oftentimes be ambiguous and fail to reflect the true nature of the work". This is particularly true for generic titles such as *Arbeiterin* [*worker* f.], *Bedienerin* [*servant f.*], *Praktikant* [*intern*], *Lehrling* [*apprentice*], which do not express sufficient information about the work involved.

Given these challenges, our future research will focus on grouping job advertisements by sector and exploring methods to predict job titles based on the job description, as demonstrated by (Huynh et al., 2019). This approach could help address the ambiguity and lack of specificity often found in historical job titles.

Limitations

One limitation encountered during our work was a lack of definition of what a job title is. While in most ads, the titles are explicit, there is a not-negligible number of ads containing just words like *Mädchen* [*girl*] or *Mann* [*man*]. The lack of definition caused minor inconsistencies among annotators which may slightly skew the results. It is also important to note, that the two machine learning approaches, namely NER and text generation, are not trained on the same task. The NER approach is trained on token classification and predicts the exact location of a job position. Whereas the text generation approach rewrites the original text to only consist of job positions. While this difference is not of major significance for our specific task, this can be relevant to consider for potential use cases. Another limitation is found during the lemmatization step. Only a very limited number of resources for lemmatization of historical german texts are currently available and are mostly rule based. As such, job positions may not be properly normalized for evaluation, when they are not covered by the utilized set of rules.

Acknowledgments

We 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. We are also thankful to Vladimir Matlach for his consultations and helpful insights. This work was supported by the FWF under grant number P 35783.

Code and Data Availability

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

References

Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., & Picariello, A. (2015). Classification of Web Job Advertisements: A Case Study. 144–151.

Bandara, R., Gunasekara, H., Peiris, W., Wijekoon, W., De Silva, T., Hewawalpita, S., & Rathnayake,

H. (2021). Information extraction from Sri Lankan job advertisements via rule-based approach.

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2017). Using machine learning for labour market intelligence. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10, 330–342.

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Classifying online Job Advertisements through Machine Learning. Future Generation Computer Systems, 86, 319–328. https://doi.org/10.1016/j.future.2018.03.035

Colace, F., De Santo, M., Lombardi, M., Mercorio, F., Mezzanzanica, M., & Pascale, F. (2019). Towards labour market intelligence through topic modelling.

Collins, M., & Singer, Y. (2002). Unsupervised Models for Named Entity Classification. Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC.

Cucerzan, S., & Yarowsky, D. (2001). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Fernández-Reyes, F. C., & Shinde, S. (2019). CV Retrieval System based on job description matching using hybrid word embeddings. Computer Speech & Language, 56, 73–79. https://doi.org/10.1016/j.csl.2019.01.003

Gnehm, A.-S., Bühlmann, E., Buchs, H., & Clematide, S. (2022). Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads. Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), 14–24. https://doi.org/10.5167/uzh-230653

Gnehm, A.-S., & Clematide, S. (2020). Text Zoning and Classification for Job Advertisements in German, French and English. In D. Bamman, D. Hovy, D. Jurgens, B. O'Connor, & S. Volkova (Eds.), Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science (pp. 83–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlpcss-1.10

Grüger, J., & Dr. Schneider, G. (2019). Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements (p. 233). https://doi.org/10.5220/0008068202260233

Guo, S., Alamudun, F., & Hammond, T. (2016). RésuMatcher: A personalized résumé-job matching system. Expert Systems with Applications, 60, 169–182. https://doi.org/10.1016/j.eswa.2016.04.013

Hobbs, J. R., & Riloff, E. (2010). Information Extraction. Handbook of Natural Language Processing, 2.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Huynh, T., Nguyen, K., Nguyen, N., & Nguyen, A. (2019). Job Prediction: From Deep Neural Network Models to Applications.

I. Rahhal, K. M. Carley, I. Kassou, & M. Ghogho. (2023). Two Stage Job Title Identification System for Online Job Advertisements. IEEE Access, 11, 19073–19092. https://doi.org/10.1109/ACCESS.2023.3247866

Grover, C., Givon, S., Tobin, R., & Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 1343–1346.

International Institute of Social History. (2023). *History of Work—HISCO*. International Institute of Social History. https://iisg.amsterdam/en/data/data-websites/history-of-work

Kelleher, J. D. (2019). *Deep learning*. MIT press.

Labusch, K., Zu, S., Kulturbesitz, B., Neudecker, C., & Zellhöfer, D. (2019, October). *BERT for Named Entity Recognition in Contemporary and Historical German*.

J. Malinowski, T. Keim, O. Wendt, & T. Weitzel. (2006). Matching People and Jobs: A Bilateral Recommendation Approach. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 6, 137c–137c. https://doi.org/10.1109/HICSS.2006.266

Jurish, B. (2012). Finite-state Canonicalization Techniques for Historical German. [Universität Potsdam]. urn:nbn:de:kobv:517-opus-55789

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. https://arxiv.org/abs/2401.10825

Leeuwen, M. H. D. van, Edvisson, S., Maas, I., & Miles, A. (2002). HISCO: Historical International Standard Classification of Occupations. https://iisg.amsterdam/en/data/data-websites/history-of-work

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics. Doklady, 10, 707–710.

Litecky, C., Aken, A., Ahmad, A., & Nelson, H. (2010). Mining for Computing Jobs. Software, IEEE, 27, 78–85. https://doi.org/10.1109/MS.2009.150

Malherbe, E., Cataldi, M., & Ballatore, A. (2015). Bringing Order to the Job Market: Efficient Job Offer Categorization in E-Recruitment.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text Annotation Tool for Human. https://github.com/doccano/doccano

Neculoiu, P., Versteegh, M., & Rotaru, M. (2016). Learning Text Similarity with Siamese Recurrent Networks. https://doi.org/10.18653/v1/W16-1617

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

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

S. Chala, S. Harrison, & M. Fathi. (2017). Knowledge extraction from online vacancies for effective job matching. 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 1–4. https://doi.org/10.1109/CCECE.2017.7946793

Sayfullina, L., Malmi, E., & Kannala, J. (2018). Learning Representations for Soft Skill Matching. In W. M. P. van der Aalst, V. Batagelj, G. Glavaš, D. I. Ignatov, M. Khachay, S. O. Kuznetsov, O. Koltsova, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, & A. V. Savchenko (Eds.), Analysis of Images, Social Networks and Texts (pp. 141–152). Springer International Publishing.

SpaCy. (n.d.). *Model Architectures*. https://spacy.io/api/architectures#Tok2Vec [3.10.2024]

SpaCy. (n.d.). *Model Architectures*. https://spacy.io/api/architectures#TransformerModel [3.10.2024]

Tavan, E., & Najafi, M. (2022). MarSan at SemEval-2022 Task 11: Multilingual complex named entity recognition using T5 and transformer encoder. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, & S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 1639–1647). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.226

Ternikov, A. (2022). Soft and hard skills identification: Insights from IT job advertisements in the CIS region. PeerJ Computer Science, 8, e946.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 142–147. https://aclanthology.org/W03-0419

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. https://arxiv.org/abs/2304.10428

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, *5*. https://doi.org/10.3389/fdigh.2018.00002

Wowczko, I. (2015). Skills and Vacancy Analysis with Data Mining Techniques. Informatics, 2, 31–49. https://doi.org/10.3390/informatics2040031

Zhu, Y., Javed, F., & Ozturk, O. (2017). Document embedding strategies for job title classification. The Thirtieth International Flairs Conference.