A Human Perspective to AI-based Candidate Screening

Laura Vásquez-Rodríguez[1], **Bertrand Audrin**[2], **Samuel Michel**[1],
Samuele Galli[3], **Julneth Rogenhofer**[2], **Jacopo Negro Cusa**[3], **Lonneke van der Plas**[1,4,5]
[1]Idiap Research Institute, Switzerland
[2]EHL Hospitality Business School, HES-SO, University of Applied Sciences and Arts Western Switzerland, Switzerland
[3]Arca24.com SA, Switzerland
[4]Institute of Linguistics and Language Technology, University of Malta, Malta
[5]Università della Svizzera Italiana, Switzerland
laura.vasquez@idiap.ch, bertrand.audrin@ehl.ch, samuel.michel@idiap.ch
samuele.galli92@gmail.com, julneth.rogenhofer@ehl.ch
j.negrocusa@gmail.com, lonneke.vanderplas@usi.ch

Abstract

Skill extraction is at the core of algorithmic hiring. It is based on identifying terms commonly found in both targets (i.e., resumes and job offers), aiming at identifying a "match" or correspondence between both. This paper focuses on skill extraction from resumes, as opposed to job offers, and considers this task both from the Human Resource Management (HRM) and AI points of view. We discuss challenges identified by both fields and explain how collaboration is instrumental for a successful digital transformation of HRM. We argue that annotation efforts are an ideal example of where collaboration between both fields is needed and present an annotation effort on 46 resumes with 41 trained annotators, resulting in a total of 116 annotations. We analyze the skills extracted by multiple different systems and compare those to the skills selected by the annotators, and find that the skills extracted differ a lot in terms of length and semantic content. The skills extracted with conversational Large Language Models (LLMs) tend to be very long and detailed, other systems are very concise, whereas humans are in the middle. In terms of semantic similarity, conversational LLMs are closer to human outputs than other systems. Our analysis proposes a different perspective to understand the well-studied, but still unsolved skill extraction task. Finally, we provide recommendations for the skill extraction task that aligns with both HR and computational perspectives.[1]

Keywords: Skill Extraction, Human Resource Management, Natural Language Processing, AI, Candidate Screening.

[1]We will release our code on GitHub: https://github.com/idiap/human_skill_extraction

1. Introduction

The role of technology in Human Resource Management (HRM) has always been important (Marler & Fisher, 2013), but the recent developments of Artificial Intelligence (AI) and Natural Language Processing (NLP) have offered new promises for HRM practices (Cheng & Hackett, 2021). Under the label of "HR Tech" (Nyathani, 2023), many traditional HRM practices have taken an AI-turn throughout the whole employee life cycle, managing activities such as employee onboarding, performance management, and learning (Prikshat et al., 2023). In this paper, we focus on the recruiting stage, where companies often receive hundreds of applications for a single position (Davis & Samaniego de la Parra, 2024). Automation thus appears as the only way to deal with such high volumes of applicants, overlooking the human aspect and often limiting the contribution of HR professionals.

In the field of NLP, we observe several attempts at automating candidate screening. Nowadays, skill extraction is the preferred and straightforward approach for automatically matching potential candidates with job offers (Shi et al., 2020). Skills, knowledge, certifications, and experience are identified in both texts, and the similarity between both is used to approximate the suitability of a given candidate for a specific job. The extraction of relevant terms can be performed using rule-based systems, as well as machine learning approaches using semantic similarity (Gugnani & Misra, 2020), and, recently, conversational LLMs (N. Li et al., 2023). Such skill extraction methods often limit the role of HR specialists and focus on the extraction from job offers, because this data is more easily accessible.

The main objective of our study is to identify the challenges associated with the skill extraction task, in particular from resumes. We discuss this task both

URI: https://hdl.handle.net/10125/109427
978-0-9981331-8-8

from the HR and computational perspective,[2] and show how both perspectives are important especially when it comes to the annotation of resumes. We construct a dataset for skill extraction by running an annotation task on 46 resumes with 41 trained annotators, resulting in a total of 116 annotations. We provide a comparison of the nature of the skills extracted by humans and machines from a set of resumes we collected, thereby showing 1) how the skills extracted by different systems vary in terms of length and semantic content, and 2) how they differ from skills extracted by humans. These analyses not only support the understanding of the challenges of a well-studied but still unsolved task but, they also suggest that a shift to a Human-AI teaming perspective would be beneficial. We finally provide a set of recommendations regarding skill extraction which aims to bring together HR researchers, AI researchers, and HR tech professionals.

2. Background

In this section, we present the skill extraction task from different perspectives. First, we discuss the HR perspective and analyze the process of resume screening (Section 2.1). Next, we discuss the skill extraction task from a computational perspective (Section 2.2). Finally, we explain how the annotation of resumes lies at the intersection of both fields (Section 2.3).

2.1. Resume Screening from an HR Perspective

On average, recruiters spend 10 seconds on a resume (Hangartner et al., 2021). Their focus is thus very likely to be on high-level aspects of the resume rather than on details, which might go unnoticed. This short time frame also suggests that the recruiters' experience and perception play a major role and that elements such as phrasing or visual aspects can have a strong influence on resume screening (Jan Ketil Arnulf & Larssen, 2010).

The whole challenge of an initial resume evaluation lies in matching a position with a profile (Kristof-Brown, 2000), based on limited information. In some instances - for conventional positions such as "software developer" - the specific technical skills that are required can be easily mapped out and identified in resumes (such as "Python" or "Scrum Master"). In other instances, some position requirements in terms of skills might be harder to identify and are sometimes even open to interpretations, especially if soft skills are the main inputs. Soft skills are a combination of personality traits, goals, motivations, and preferences that tend to be increasingly valued in the workplace (Heckman & Kautz, 2012). These tend to be expressed in multiple ways according to the experience and background of the candidate. Hence, the detection and standardization of soft skills are also more challenging aspects (Khaouja et al., 2019). Another challenging aspect is related to the job offers themselves, as many required skills are not explicit, and in many companies, job offers are generic and not tailored to the specific requirements of a job (Gugnani & Misra, 2020). In contrast, resumes are highly heterogeneous as they strongly depend on the candidates and their background (industry, culture, language), which will affect how ideas, skills, and the overall resume will be presented (Sajid et al., 2022).

Another perspective to keep in mind is that of candidates who face the challenge of having to convince HR professionals as well as algorithms, with potential tensions between both. On the one hand, presenting a convincing resume to HR professionals requires visual cues and formatting; on the other hand, presenting a convincing resume to an "algorithmic audience" requires focusing on appealing keywords tailored for a specific job posting. Candidates can use elaborate templates to make their resumes stand out to the human eye, while also relying on conversational LLMs to craft their resume for a specific job offer. This results in a tension between customization and standardization in the content and styles of resumes, making pre-selection even harder.

2.2. Skill Extraction from a Computational Perspective

Algorithmic hiring has focused on skill extraction and has become a prominent field of study (Senger et al., 2024; Zhang et al., 2023). Previous work encompasses both hard skills (Goyal et al., 2023; Zhang et al., 2024) and, to a lesser extent, also soft skills (Sayfullina et al., 2018; Zhang, Jensen, et al., 2022). An even smaller number of works also focus on the precise match of candidates and job offers (Guo et al., 2016).

Often, the skill extraction task is tackled by using a taxonomy of terms to search (Cenikj et al., 2021) for co-incidences in the text (e.g., rule-based system). A taxonomy is a collection of terms that interconnect occupations with skills for a specific domain as shown in Figure 1. In this scenario, a skill is extracted when words or phrases in the taxonomy are identified in the job description or resume. Due to its strict adherence to matches between words and phrases, these rule-based systems are conservative in that they extract only a limited number of often quite precise skills. This behavior also explains why these systems often overlook

[2] By using existing computational methods to shed light on the task of skill extraction (Padmanabhan et al., 2022).

certain skills, that might have been described in a way that deviates from the terms and phrases described in the taxonomy. Nevertheless, rule-based systems are effective and easy to interpret also thanks to the direct link to an overarching taxonomy (Rudin, 2019).

As an alternative, machine-learning methods have been incrementally introduced, where models can learn to extract terms that have not been seen before. The advantage of these methods is that their ability to learn can help these models to generalize, i.e., to find different ways of expressing the same skill. We can roughly classify previous work into models that rely on annotated data to fine-tune pre-trained language models to identify skills in text (Green et al., 2022; Zhang, Jensen, et al., 2022) and those that do not rely on annotated data but use conversational LLMs to generate skills when given a prompt containing instruction and either a job description or resume as input (Magron et al., 2024; Nguyen et al., 2024). The main drawback of the former is the fact that these models depend on annotated data for any given language or domain of interest. The main drawbacks of the latter are relatively high computational cost and limited control over the output. Previous work has demonstrated that the output of conversational LLMs for skill extraction results in 36% disagreement with humans (Nguyen et al., 2024). We elaborate on the annotation challenge in Section 2.3.

Multiple limitations can be found in previous work on automatic skill extraction (Goyal et al., 2023; Zhang, Jensen, & Plank, 2022; Zhang et al., 2024). Firstly, the focus has been mainly on extracting skills from job descriptions (Goyal et al., 2023; Zhang et al., 2024) and largely ignored the task of extracting skills from resumes, which is an indispensable element of any Applicant Tracking System (ATS). One of the reasons for this discrepancy is practical. Resumes contain personal data and are therefore very hard to obtain and release as research data, nowadays often required for publishing research papers. Moreover, using commercial LLMs for extracting skills from resumes raises severe privacy concerns. However, resumes are very different from job descriptions, and a system that performs well on skill extraction from job descriptions is not guaranteed to work well on resumes. Resumes often use complicated layouts, for reasons discussed earlier, that are hard to parse automatically, among other issues.

Previous work has overlooked the fact that businesses often have clients from all parts of the world and many different expertise areas. Annotated data is not available for all languages and domains, and model transfer may also fail. The definitions and identification of what a skill is might also vary. For example, in the medical domain certifications are important while in business domains soft skills are essential. To alleviate issues related to data scarcity, datasets are synthetically generated datasets using conversational LLMs (Decorte et al., 2023; Magron et al., 2024). Such methods could potentially also mitigate issues regarding privacy. However, the relevance of these methods for a skill identification task that can be considered as subjective - to an extent - is still questionable (Z. Li et al., 2023).

The difference between manual resume screening and automatic skill extraction is quite large. First, computational systems are not trained on screening resumes based on visual elements. Second, the skill extraction task is almost always decoupled from the actual job offer in computational methods. In the next subsection, we will explain what consequences this has for the manual annotation of resumes.

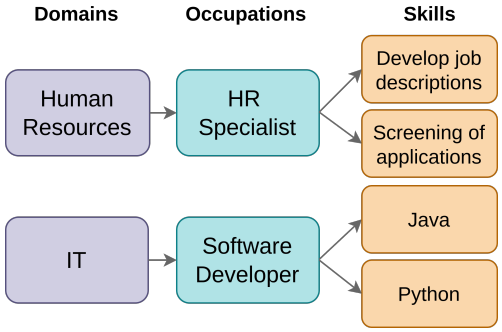


Figure 1: Example of a skills taxonomy.

2.3. AI and HR Intersection: Annotations

Supervised machine-learning approaches for skill extraction require the collection of human annotations (Tamburri et al., 2020). Traditionally, researchers in the AI fieldwork with annotators from the HR domain to perform this task (Zhang, Jensen, et al., 2022). HR professionals then create annotations independently, according to defined guidelines, and often, annotations are rarely published (Senger et al., 2024). In this project, we created annotations and annotation guidelines in a collaboration between HR researchers, AI researchers, and HR tech professionals to include all three perspectives. We have learned several important aspects regarding the annotation process that we discuss below.

Overall, the manual annotation of job offers and especially resumes is an extremely subjective task. First, it is important to note that resumes represent a very specific form of text that is highly fragmented (resumes often feature only a few full sentences but a high

number of lists and bullet points) and codified (specific information is expected in a resume, such as education, experience). However, candidates benefit from a high level of autonomy within these constraints, which leads to substantial variability between different resumes, and with it complexity when annotating their content.

The context of the job application (e.g., type of position, language) process also plays an important role in shaping resumes. In fact, the entire purpose of the resume is for a candidate to showcase that their profile matches with the position that they are applying for. A resume is thus expected to "target" a job offer. This is critical in two respects: 1) how it influences the resume itself, and 2) how it influences the annotation process. The context actively shapes the content of a resume: people applying for software developer jobs are more likely to emphasize the projects they have worked on, and the programming languages that they mastered, whereas people applying to human resource managers jobs are more likely to focus on their tasks and responsibilities and the results. The context also shapes the annotation itself. Indeed, HR professionals always compare resumes with a job description and a job specification: there is always a point of reference that can be used to contextualize the experience and the skills that can be found in a resume. An out-of-context resume is thus more difficult to analyze and its annotation requires more expertise and stricter guidelines.

In that respect, annotators require a high level of expertise in the domain of the resumes that they are annotating (e.g. computer science, finance, human resource management) as they are not able to rely on existing job offers to assist them in their analysis. This expertise also needs to be complemented with a clear set of guidelines that clarify what to annotate and how. This is particularly critical in skill annotation as different forms of competence can be featured in a resume (e.g., theoretical knowledge, practical experience, abilities, etc.), which can lead to confusion in annotations. Having clear categories of skills with examples for different contexts thus helps reduce uncertainty and increase reliability throughout the annotations.

Given these conditions, annotating skill datasets for AI represents one of the biggest challenges in developing HR tech solutions, and considerable time should be reserved for this phase.

3. Methods

This section describes the data collection process, with the gathering of human annotations in the HR domain (Section 3.1). Next, we compare the human annotations with those generated automatically, using the same corpus and rule-based and AI-based systems (Section 3.2). In addition, we performed a comparative analysis (Section 3.3) regarding the length and semantic content of our annotations to better understand the challenges of the annotation task.

Skills	Mean p/resume	Mean p/resume (unique)	Total	Total (Unique)
Human	30.522	24.478	1404	1126
LLM	52.391	49.652	2410	2284
Rule-based	16.478	12.848	758	591
Supervised	37.022	36.022	1703	1657

Table 1: Statistics for skills extracted.

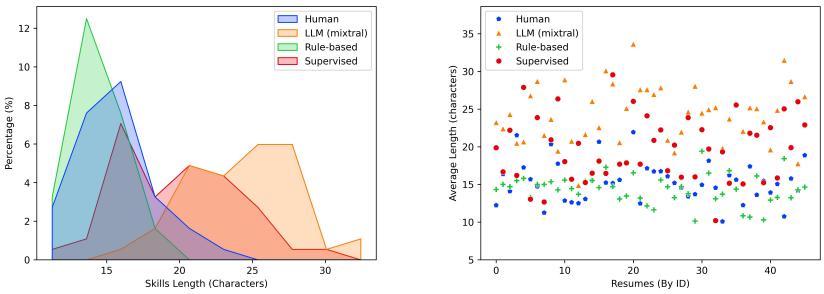
3.1. Data Collection: Human

We collected a resumes-based dataset suitable for the skill extraction task. To achieve this goal, we organized a workshop with students of a graduate-level Talent Management course in the hospitality domain. The main task was to craft their resumes to collectively gather further annotations.[3] Students annotated each other's resumes and we ensured that annotators were all specialists in the domains they were annotating. As we previously mentioned, it is of high importance when annotating resumes without a specific job description to match against, that annotators are domain specialists. The annotators were divided into groups of 3-4 people and individually, they annotated each one resume. Each resume was divided into paragraphs (if possible), annotating in each case, the corresponding label into three categories: Hard Skills/Certifications, Soft Skills, and Occupations/Positions. In total, we obtained the data consent to annotate 46 resumes (out of 60 resumes). We had a workshop participation of 41 students, who achieved a total of 116 annotated texts, given that each CV would be annotated multiple times. Also, due to the variability in the annotators' speed, some students managed to annotate more texts, completing a total of 46 CVs, after aggregating all the annotations. We report the statistics of the skills extracted in Table 2 and refer to this corpus as the *EHL workshop dataset*.

3.2. System Selection

To compare with human annotations, we first selected a **rule-based system** which is often used for skill extraction and is usually straight-forward and low-cost. A rule-based system is a program that will search keywords based on a defined set of rules or taxonomy. For our use case, we have an $in\text{-}house$

[3]The number of resumes was determined by the students who formally agreed to share their resumes for the workshop.



(a) Skill Length Distribution (b) Skill Length per Resume

Figure 2: On the left, we present the distribution (in percentages, *y-axis*) of the skills given their length and similarity (*x-axis*. On the right, we present the average length of skills and similarity scores (*y-axis*) per resume (*y-axis*).

	Mean per resume	**Std. per resume**	**Total values**
Words	465.391	181.214	21,408
Annotators	2.522	0.836	41

Table 2: Statistics for the *EHL workshop dataset* of 46 resumes.

industrial taxonomy with triplets that define a domain, an occupation, and its corresponding skills as shown in Figure 1. We extracted skills in the dataset by using methods based on the open-source SkillNER tool[4] and identified the labels that matched with the text and the taxonomy based on the predefined rules.

For the AI-based methods, we selected two approaches: a **supervised model** and a **conversational LLM**. The supervised model is trained on annotated corpora, specializing the model with domain-specific knowledge, such as job offers or skill extraction concepts, and to the language of the datasets. These supervised models tend to be smaller than a conversational LLM and they do not need large amounts of examples to perform a task (i.e., using fine-tuning methods). In particular, we used conversational LLMs, which were trained on large amounts of data. They are often used as is, as retraining them would involve extensive computational resources.

3.3. Comparative Analysis

We summarize the collected annotations discussed in the previous subsections in Table 1. Next, we analyze the collected extractions as follows: an analysis of the annotations based on their surface properties (i.e., skills length) and based on their meaning (i.e., skills' semantic similarity).

For the **length analysis**, we calculate the average length of the extractions within each resume, independently between systems and humans. Given this information, first, we analyze the distribution of the skill lengths aggregating all resumes as in Figure 2a for each data collection (e.g., human, supervised). We normalized all the observations using the *percent* statistic,[5] so labels from each source will represent a portion within totality (i.e., 100%) of the observations. Second, we look individually at the average length differences per resume as in Figure 2b. This shows the individual contribution of each resume to the length measurement.

For the **semantic analysis**, we proposed the comparison of the meaning of the extracted skills between the following pairs: human vs. LLM, human vs. rule-based, and human vs supervised. This experiment aims to reflect similarities between skills collected for a given resume by humans and the different systems. We calculated the semantic similarity between resumes using the cosine similarity metric.[6] This metric represents documents (i.e., groups of skills) as abstract representations where the distance between them reflects how close they are in meaning. Similar skills would have a value near 1, while dissimilar labels would show values close to 0. Similarly to the length analysis, we analyze the distribution (by percentage) of the similarity values from the systems' outputs. Also, we include an analysis per resume to understand the individual contribution to the overall scores. For both analyses, it is important to note that it is very unlikely to have a 1:1 correspondence between all skills. Therefore, we have grouped the annotations within each system and resume as reported in the length analysis. We will give more details in Section 6.

[4]https://github.com/AnasAito/SkillNER

[5]https://seaborn.pydata.org/generated/seaborn.histplot.html

[6]https://www.sbert.net/docs/package_reference/util.html?highlight=cos_sim#sentence_transformers.util.cos_sim

Source	Annotations
Human	negotiation, market research, project planning, language skill- english, microsoft 365, training skills, tlb management, marketing and strategy, culinary certificate, interview experience
LLM	forbes standards, contract negotiation, coaching, multitasking, brand image and communication strategy, menu planning, budget & financial management, carrying out feasibility studies of new projects/refurbishment projects, design, managing large groups
Rule-based	hotel led, contractor, management graduated, mathematics, sports, prepared materials, digital strategy, business schools, sales consultant, operational processes provided support
Supervised	maintained consistent communication with customers, nexus, overseeing, led the role of a project manager, my vision, degree, and maintained real estate project tracking and portfolio, systems, with market research, the restaurants daily operations, sustainability

Table 3: Manual and automatic annotations from our dataset. We selected 10 random annotations from each model.

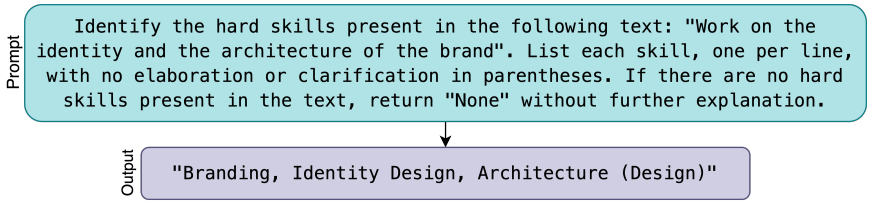


Figure 3: Prompt example for annotating resumes with LLMs.

4. Experimental Section

In this section, we discuss the technical details of our experiments. We explain the preprocessing steps of our dataset (Subsection 4.1) and include the implementation details of our models and the comparative analysis (Subsection 4.2).

4.1. Data Processing

We aggregated all the annotations from each annotator of the *EHL workshop dataset*. As a result, each resume would have a single list of labels (unique / non-unique) as shown in Table 3 (Human). We also removed inconsistencies such as different characters for separating skills. Finally, for the scope of this paper, we have selected only the analysis of hard skills as soft skills are difficult to detect by automatic methods. In future work, we will include the detection of soft skills in our tools and taxonomies.

4.2. Models and Analysis

For the **conversational LLM**, we specified the model task, with a set of instructions, explicitly formatting the input text as shown in Figure 3. For the automatic annotation task, we carried out the analysis using the open-source model *Mixtral 8x7B*[7] that can be run locally given that resumes handle personal information, in contrast to online commercial models that upload the data provided in each query (Nguyen et al., 2024). We use these models as a means to automatically extract skills from the *EHL workshop dataset*.

For the **supervised baseline**, we aimed to compare a state-of-the-art supervised model. Hence, we selected the model ESCOXLM−R, which has been previously trained on the ESCO taxonomy (i.e., mainly skills/competences and occupations), and samples of annotated job offers.[8] This model encompasses two specializations, detection of skills[9] and knowledge.[10] We used these models to annotate the text in our dataset with no further training.

Concerning the implementation details of the semantic analysis, we calculated the similarity between text representations using the SBERT (Reimers & Gurevych, 2019) model,[11] which reports the highest average performance for the semantic similarity task[12] For the length analysis, the implementation is straight-forward, as we report the total of characters in a skill and then average for all the labels in the resume.

5. Results

We present our length experiments in Figure 2. From Figure 2a, we can observe that most human annotations have an approximate length of 15 characters, while skills generated by conversational LLMs are generally more verbose, with varying lengths between 15 and 35. In contrast, the rule-based is shown to be more concise, but still closer to the length of human annotations. The

[7] https://huggingface.co/TheBloke/Mixtral-8x7B-Instruct-v0.1-GGUF

[8] https://esco.ec.europa.eu/en/about-esco/what-esco

[9] https://huggingface.co/jjzha/escoxlmr_skill_extraction

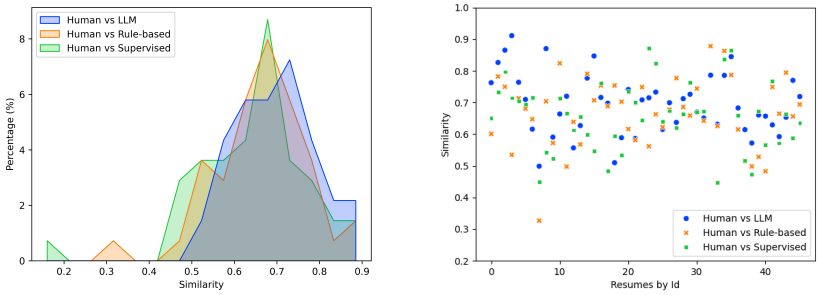
[10] https://huggingface.co/jjzha/escoxlmr_knowledge_extraction

[11] *all-mpnet-base-v2*

[12] https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#original-models

Human	Rule-based	LLM	Supervised
data collection and analysis	data analysis	determined the origin of lots/items through desk research, data, and sources analysis	data analysis
budget management	budget	estimated tender budget	Estimated tender budget
interview experience	—	conducted facetoface interviews	Conducted face-to-face interviews with airport users
customer service, office	front desk	receiving phone calls, customer service, familiarity with customer checkin and checkout procedures	provide customers with thoughtful service, reply to, serve every guest warmly and thoughtfully, the phone, with customers

Table 4: Common annotations between humans and our proposed models.



(a) Similarity frequency

(b) Similarity by resume

Figure 4: On the left, we present the distribution (*y-axis*, percentage) of the observations within their semantic similarity scores (*x-axis*). On the right, we show the semantic similarity values (*y-axis*) distributed across resumes (*x-axis*).

supervised method results in extractions that are more wordy than human annotations, but it is more concise than the conversational LLM. From the analysis per resume (Figure 2b), it is clear that conversational LLMs produce the longest extractions, with some exceptions from the supervised method.

Next, we perform a semantic analysis in Figure 4. In Figure 4a, we plot the semantic similarity of outputs of the three systems in comparison to human extractions aggregated over all resumes. We can see that most of the documents are somewhat related, with values classified between 0.5 and 0.8. However, there are no examples with values between 0.9 and 1, showing that there are no exact matches. Also, we present the semantic similarity per resume in Figure 4b. We compare the similarity of the human annotations against all the systems. Once again, no system produces outputs that are highly similar to human output, but there is considerable semantic overlap because the similarity between them still ranges between 0.5 and 0.8, with the conversational LLM showing the highest similarity with human content.

Finally, to better understand the quality of our quantitative analysis, we provide a list of extractions, both from humans and systems. In Table 3, we show a list of 10 randomly selected skill extractions for each model. We argue that the random selection of samples helps to understand the variability of the labels (in size and semantics) of each model. Furthermore, in Table 4, we manually mapped common concepts from 5 random resumes to point out some distinctions and similarities between the different approaches.

6. Discussion

First, we comment on our comparative analysis. Rule-based systems are shown to be more standardized, as they adhere to a predefined taxonomy where skills have a maximum length. In contrast, conversational LLMs are generative models (i.e., they output text based on previously seen texts) that output not only explicitly mentioned skills but may also infer content from the text. Supervised models, tend to stay closer to the actual task they have been explicitly trained to do, but will be able to abstract away from literal mentions of a particular skill as given in a taxonomy. The aim of training such models is for them to be able to generalize well to previously unseen skills. It is difficult to conclude the superiority of one system over the other from these analyses, and this is also not the aim of this paper. These analyses bring to light some characteristics of these system outputs: In the context of this study, conversational LLMs show the highest overlap in semantic content, but they are also extremely

wordy, rule-based keyword-matching systems produce very short outputs that are linked to the items found in a taxonomy that seem to cover the semantic content a bit less well, and supervised are somewhere in between.

Concerning our analyses, we highlight that these results should be considered with caution, as they result from a small dataset of 46 observations. We acknowledge that their generalization can be limited and further investigations with larger datasets are needed to formulate more decisive statements. Nevertheless, the results from our experiments can help create awareness of the variability in output across different model architectures and humans beyond pure numbers. Concerning the conversational LLM-based skill extraction, we present the results for a specific LLM - *Mixtral-8x7B model* - with our customized, single prompt. These models are highly variable and non-deterministic, so results may vary according to the architecture and prompts. Prompts could include a length limitation, for example. Although the definition of a taxonomy in the context of the LLM could mitigate the variable nature of the skill extraction, these could be extremely large,[13] representing non-trivial challenges for its implementation.

Finally, we discuss our learnings from the analyses. As mentioned before, from the human perspective, the skill extraction task can be subjective due to the differences in people's backgrounds and domains. On the computational side, we can confirm how methods annotating the same resumes can vary greatly as well. Therefore, the development of clear and well-defined guidelines is imperative to ensure more objectivity in the task. While annotation guidelines have been released in the past (Senger et al., 2024), following an elaboration process that is specific to the hiring and recruitment context and involves HR professionals, is instrumental to creating clear documentation and ensuring a more structured process. Also, while comparing the different computational approaches, the human element should not be forgotten. A close inspection of system output and careful consideration of elements such as computational resources needed, privacy concerns, and explainability should be taken into account in addition to performance figures on the task. We conclude that this collaboration is essential for a fair, equal, and effective approach to candidate screening.

7. Recommendations and Future Work

In this section, we present our recommendations to encourage closing the existing gap between AI researchers, HR researchers, and HR tech professionals.

[13] https://esco.ec.europa.eu/en/classification/skill-main

Interdisciplinary collaboration and partnership with industry From our collaboration efforts in our current HR tech project, we noted the importance for NLP researchers to work closely with HR researchers so that both sides benefit from each other's expertise. We also experienced that it took some time for the interdisciplinary team to find a common language and develop a mutual understanding of the task at hand and the contributions from both sides. This has been particularly instrumental in the annotation task to ensure that the output meets the requirements from both NLP and HR perspectives. In the same vein, partnering with industry is also fundamental, not only to understand the practical application of a given solution (e.g., skill extraction) but also for the implications and benefits to the business, which could include the adaptation to existing taxonomies, limited computational resources and compatibility with existing solutions. In the context of this project, this has led us to focus on how to meet objectives from an industry perspective and also in terms of academic research: developing experimentation and novelty to thrive, while at the same time keeping an eye on cost, time-effectiveness, and long-term maintenance of systems.

Another important aspect is the trust in selection algorithms (Groß, 2021). While there is a strong debate among practitioners on whether algorithms can be trusted or not for selection decisions, our results highlight a certain degree of variability between systems and human annotations in terms of length and meaning. Also, we explained that the output from certain methods (rule-based) is easier to explain than others. These observations are linked to well-known challenges of explainability, predictability, precision, and overall trustworthiness of AI-based recruitment systems. In that respect, building robust and explainable systems seems to be a priority to increase the trust of users and the adoption of the technologies, also in view of AI regulations.

Limitations and Future Work We have proposed the comparison of multiple models as a means to understand the variability of different sources. Nevertheless, it is important to state their limitations. For the supervised setting, we chose publicly available state-of-the-art models, however, these are only trained on job offers as resumes are not freely available. For the rule-based, we are limited by the predefined taxonomy of skill concepts, thus, the performance will depend on the quality and nature of the taxonomy. Concerning the conversational LLMs, we chose a zero-shot approach to generate extractions from our dataset. However, it would also be beneficial to experiment with few-shot

scenarios and in-context learning so that the models can see practical examples and also, it is possible to provide additional feedback to improve the quality of the output.

In future work, we would like to extend the number of annotated datasets to include resumes from various domains and languages. This will improve the generalization abilities of the models to more scenarios. With respect to the conversational LLMs, these are sensitive to minimal changes in the selected prompts. By using multiple prompting techniques and models, we could arrive at a deeper understanding of their performance. This includes the limitation on the length of candidates' skills and the introduction of external knowledge from taxonomies, or other domain knowledge. In particular, we would like to include soft skills, which are fundamental in the workplace and their identification is one of the main goals of our project. However, in order to have a fair comparison between the different approaches – which is our objective in this paper – we would need a taxonomy that specifically focuses on soft skills as well. Such a taxonomy, tailored to our needs, is currently not available, so we decided to leave soft skills out of this specific work while hoping to include them in future research.

8. Conclusion

In this paper, we discussed the different perspectives of HR and AI on the skill extraction task. We focused on resumes, in contrast to previous work, and collected 46 annotated texts. These allowed us to perform comparative analyses of the output generated by three different computational methods on the task of automatic skill extraction. We show the difference in nature of what these computational systems produce. We also discuss the pros and cons of these approaches from several angles focusing on the human aspects fed by interactions with HR professionals and HR tech professionals, such as privacy, trustworthiness, and efficiency. We expect that our research sheds light on the variations between systems and how they differ from human annotations, emphasizing the role of this often neglected actor in a field that tends to rely more and more on algorithmic hiring.

9. Acknowledgements

We would like to thank Alexandre Nanchen for his feedback on this paper. Finally, we gratefully acknowledge the support from Innosuisse (grant 104.069 IP-ICT).

References

Cenikj, G., Vitanova, B., & Eftimov, T. (2021). Skills named-entity recognition for creating a skill inventory of today's workplace. *2021 IEEE International Conference on Big Data (Big Data)*, 4561–4565.

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in hrm: Definition, theory, and practice. *Human Resource Management Review*, *31*(1), 100698.

Davis, S. J., & Samaniego de la Parra, B. (2024). *Application flows* (Working Paper No. 32320). National Bureau of Economic Research.

Decorte, J.-J., Verlinden, S., Hautte, J. V., Deleu, J., Develder, C., & Demeester, T. (2023). Extreme multi-label skill extraction training using large language models.

Goyal, N., Kalra, J., Sharma, C., Mutharaju, R., Sachdeva, N., & Kumaraguru, P. (2023). JobXMLC: EXtreme multi-label classification of job skills with graph neural networks. *Findings of the Association for Computational Linguistics: EACL 2023*, 2181–2191.

Green, T., Maynard, D., & Lin, C. (2022). Development of a benchmark corpus to support entity recognition in job descriptions. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1201–1208.

Groß, M. (2021). Yes, AI Can: The Artificial Intelligence Gold Rush Between Optimistic HR Software Providers, Skeptical HR Managers, and Corporate Ethical Virtues. In *AI for the Good* (pp. 191–225). Springer.

Gugnani, A., & Misra, H. (2020). Implicit skills extraction using document embedding and its use in job recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(08), 13286–13293.

Guo, S., Alamudun, F., & Hammond, T. (2016). Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, *60*, 169–182.

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms, 572–576.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464.

Jan Ketil Arnulf, L. T., & Larssen, Ø. (2010). Impression making by résumé layout: Its impact on the probability of being shortlisted. *European Journal of Work and Organizational Psychology*, *19*(2), 221–230.

Khaouja, I., Mezzour, G., Carley, K. M., & Kassou, I. (2019). Building a soft skill taxonomy from job openings. *Social Network Analysis and Mining*, *9*(1), 43.

Kristof-Brown, A. L. (2000). Perceived applicant fit: Distinguishing between recruiters' perceptions of person–job and person–organization fit. *Personnel Psychology*, *53*(3), 643–671.

Li, N., Kang, B., & Bie, T. D. (2023). Skillgpt: A restful api service for skill extraction and standardization using a large language model.

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10443–10461.

Magron, A., Dai, A., Zhang, M., Montariol, S., & Bosselut, A. (2024). JobSkape: A framework for generating synthetic job postings to enhance skill matching. *Proceedings of the First Workshop on Natural Language Processing for Human Resources 2024*, 43–58.

Marler, J. H., & Fisher, S. L. (2013). An evidence-based review of e-hrm and strategic human resource management [Emerging Issues in Theory and Research on Electronic Human Resource Management (eHRM)]. *Human Resource Management Review*, *23*(1), 18–36.

Nguyen, K., Zhang, M., Montariol, S., & Bosselut, A. (2024). Rethinking skill extraction in the job market domain using large language models. *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, 27–42.

Nyathani, R. (2023). Preparing for the future of work: How hr tech is shaping remote work. *Journal of Technology and Systems*, *5*, 60–73.

Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Machine learning in information systems research. *46*(1), iii–xix.

Prikshat, V., Islam, M., Patel, P., Malik, A., Budhwar, P., & Gupta, S. (2023). Ai-augmented hrm: Literature review and a proposed multilevel framework for future research. *Technological Forecasting and Social Change*, *193*, 122645.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sajid, H., Kanwal, J., Bhatti, S. U. R., Qureshi, S. A., Basharat, A., Hussain, S., & Khan, K. U. (2022). Resume parsing framework for e-recruitment. *16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1–8.

Sayfullina, L., Malmi, E., & Kannala, J. (2018). Learning representations for soft skill matching. *Analysis of Images, Social Networks and Texts*, 141–152.

Senger, E., Zhang, M., van der Goot, R., & Plank, B. (2024). Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, 1–15.

Shi, B., Yang, J., Guo, F., & He, Q. (2020). Salience and market-aware skill extraction for job targeting. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2871–2879.

Tamburri, D. A., Heuvel, W.-J. V. D., & Garriga, M. (2020). Dataops for societal intelligence: A data pipeline for labor market skills extraction and matching. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, 391–394.

Zhang, M., Jensen, K., Sonniks, S., & Plank, B. (2022). SkillSpan: Hard and soft skill extraction from English job postings. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4962–4984.

Zhang, M., Jensen, K. N., & Plank, B. (2022). Kompetencer: Fine-grained skill classification in Danish job postings via distant supervision and transfer learning. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 436–447.

Zhang, M., van der Goot, R., Kan, M.-Y., & Plank, B. (2024). NNOSE: Nearest neighbor occupational skill extraction. *Proceedings of the 18th Conference of the European Chapter of the ACL (Volume 1: Long Papers)*, 589–608.

Zhang, M., van der Goot, R., & Plank, B. (2023). ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, 11871–11890.