# Skill Extraction from Resumes and Job Offers across Six Languages

**Laura Vásquez-Rodríguez[1,2,*], Bertrand Audrin[3], Samuel Michel[2], Samuele Galli[4], Julneth Rogenhofer[3], Jacopo Negro Cusa[4], Lonneke van der Plas[5]**

[1]Doodle AG, Switzerland
[2]Idiap Research Institute, Switzerland
[3]EHL Hospitality Business School, HES-SO,
University of Applied Sciences and Arts Western Switzerland, Switzerland
[4]Arca24.com SA, Switzerland
[5]Università della Svizzera italiana, Switzerland

**Correspondence:** bertrand.audrin@ehl.ch, lonneke.vanderplas@usi.ch

## Abstract

We comprehensively evaluate multiple skill extraction approaches, including rule-based, semantic, and supervised methods, using resumes and job offers in English, French, German, Italian, Spanish, and Portuguese. Due to inherent privacy concerns in Human Resources (HR) data and the high cost of manual annotations, research on identifying relevant skills for the job market remains limited, often restricted to specific domains, datasets, and entity types, and is available in only a few languages. In the context of an industrial project, we have annotated 1,200 job offers and resumes across diverse domains and six languages, through a multidisciplinary collaboration among HR researchers, NLP researchers, and HR tech professionals. Our evaluation assesses the effectiveness of these systems in a multilingual, multidomain setting, capturing both standardized job offers and highly variable resumes. The results show that supervised models achieve F1 scores of up to 0.6, while rule-based methods offer better interpretability. Furthermore, we find large differences between how skills are formulated in job offers and resumes, while the latter is understudied in academic research.[1]

## 1 Introduction

In recent years, candidate selection in recruitment has increasingly relied on automated methods to identify relevant competencies in job postings and resumes (Guo et al., 2016; Boselli et al., 2018; Gan et al., 2024). Applicant Tracking Systems (ATS) primarily manage hiring processes with skill identification as a core function. However, the precision of existing methods remains insufficient to ensure a fair and unbiased recruitment process compared to human decision-making (Fabris et al., 2025). Tackling this challenge is crucial for real-world NLP applications to build long-lasting and trustworthy collaboration with industry.[2]

The task of skill extraction has been extensively studied (Senger et al., 2024; Kopparapu, 2010; Kivimäki et al., 2013; Zhao et al., 2015), using a wide variety of methods such as vector-based methods (Javed et al., 2017; Gugnani and Misra, 2020), BERT-based models (Tamburri et al., 2020; Zhang et al., 2022c, 2023) and conversational Large Language Models (LLMs) (Clavié and Soulié, 2023; Decorte et al., 2023). However, research remains narrow, focusing on specific languages, domains (e.g., IT, finance), data types (mostly job advertisements), skill categories (mostly hard skills), and extraction methods, which is very far from the actual contexts in which the HR tech industry is operating.

One of the main challenges is data scarcity, primarily due to the sensitivity of personal information in resumes. Researchers thus mainly focus on job advertisements (also referred to as job offers or job postings). While job offers are rich in career-related concepts, they strongly differ from resumes, which tend to showcase skills with more variety. Therefore, a more systematic, in-depth study is needed (Zhang et al., 2023) to address the challenges in resumes. In a practical scenario, ATS systems could not be effective when skills rely solely on job offers.

Another difficulty is the language. Linguistic differences influence how skills are presented, both in resumes and job offers. While English is sometimes considered the lingua franca for job applications (and for research on the topic), many countries primarily use their local languages, highlighting the importance of a multilingual approach. Skill extraction methods are typically designed for a monolingual setting and applied indiscriminately

[*]Work done while at Idiap Research Institute.

[1]We will publish our non-proprietary data and models on GitHub: https://github.com/idiap/multilingual_skill_extraction.

[2]We discuss our perspective towards the challenges of translating research into NLP applications in Appendix A.1.

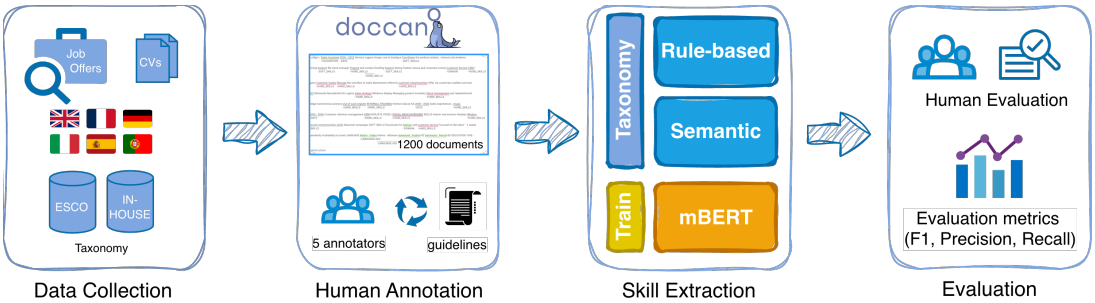


Figure 1: Overview of the end-to-end skill extraction system from *Data Collection* to the *Evaluation* stage.

across languages such as English (Zhang et al., 2022b; Bhola et al., 2020; Zhang et al., 2022a), French (Beauchemin et al., 2022), Swiss German (Gnehm et al., 2022), and Danish (Zhang et al., 2022b), without accounting for these critical differences.

A final challenge is that of explainability, which is essential to build trust. While LLMs are well known for their inference capabilities (Clavié and Soulié, 2023), they are often criticized for their lack of transparency, making it difficult to trace the source of each decision, unlike rule-based systems that rely on manually crafted taxonomies. Although rule-based approaches are not ideal due to the difficulty of keeping skill terms updated, they foster trust between systems, HR specialists, and candidates. Hybrid approaches can balance explainability and skill extraction capabilities, enhancing fairness in recruitment. These combinations can also be optimized by aligning with end-user requirements while maintaining acceptable inference speeds. Moreover, this also aligns and supports global regulations (e.g., EU Act), which urge having more transparent and fair models, especially from the deployers' side.

This study takes place within the SEM24 Innosuisse project,[3] which aims to develop skill extraction algorithms for practical applications in HR systems. This interdisciplinary initiative brings together HR specialists, HR researchers, and NLP researchers to identify skills in resumes and job offers across multiple languages and methods. Specifically, we propose a multilingual, multidomain evaluation of the skill extraction task for both job offers and resumes. For our assessment, we selected the most stable, predictable, and hardware-efficient models (Vásquez-Rodríguez et al., 2024) using handcrafted annotations. We hypothesize that exploring a broader range of methods, languages, and data types will contribute to the development of more accurate skill extraction techniques. Our key contributions are as follows:

1. A multilingual, multidomain implementation of skill extraction systems (i.e., rule-based, semantic, and supervised) for job offers and resumes;

2. A comprehensive analysis of the challenges in skill extraction across diverse scenarios, considering different languages, document types, and domains;

3. An assessment of the advantages and limitations of a deployed real-world NLP application in an industrial setting, with detailed error analysis.

## 2 Methodology

We present our task definition in Section 2.1, and further, we detail the necessary data for our experiments, including taxonomies[4] and datasets, as described in Section 2.2 and 2.3. Section 2.4 explains the annotation process and guidelines for labeling the datasets. Finally, Section 2.5 presents the selected models, which serve as the foundation for our multilingual experiments.

### 2.1 Task Definition

Our main task aims to understand the performance of the skill extraction task for both resumes and job offers, in the aforementioned scenarios, including English, French, German, Italian, Portuguese, and Spanish. The selection of these first languages was determined by the availability of the annotations.

[3] https://www.idiap.ch/en/scientific-research/projects/SEM24

[4] We refer to a taxonomy as a collection of interconnected terms that define relationships, for example, between skills and occupations. The entities that describe the taxonomy can vary between sources.

| Surface Properties → | | Total Entities | | Unique Entities | | Unique/Total | | Avg. Len | |
|---|---|---|---|---|---|---|---|---|---|
| Text ↓ | Language | Hard | Occ | Hard | Occ | Hard | Occ | Hard | Occ |
| Jobs (our annotated data) | EN | 878 | 322 | 777 | 109 | 0.88 | 0.34 | 37.53 | 22.19 |
| | FR | 1119 | 126 | 886 | 107 | 0.79 | 0.85 | 30.67 | 29.13 |
| | IT | 704 | 116 | 651 | 101 | 0.92 | 0.87 | 53.67 | 22.52 |
| | DE | 848 | 133 | 700 | 115 | 0.83 | 0.86 | 19.56 | 21.26 |
| | ES | 476 | 81 | 461 | 73 | 0.97 | 0.90 | 42.09 | 21.72 |
| | PT | 667 | 81 | 560 | 68 | 0.84 | 0.84 | 30.87 | 21.37 |
| Jobs (Green) | EN | 12573 | 2571 | 10079 | 1591 | 0.80 | 0.62 | 32.67 | 17.49 |
| Resumes (our annotated data) | EN | 4024 | 692 | 3020 | 565 | 0.75 | 0.82 | 20.01 | 24.85 |
| | FR | 2063 | 645 | 1700 | 466 | 0.82 | 0.72 | 25.76 | 21.66 |
| | IT | 1985 | 645 | 1613 | 435 | 0.81 | 0.67 | 26.77 | 21.91 |
| | DE | 2075 | 691 | 1592 | 445 | 0.77 | 0.64 | 24.35 | 17.43 |
| | ES | 1086 | 271 | 922 | 241 | 0.85 | 0.89 | 30.74 | 25.42 |
| | PT | 3312 | 729 | 2439 | 447 | 0.74 | 0.61 | 28.79 | 24.78 |

Table 1: We report the number of total and unique entities and its ratio, and average character length of hard skills, and occupations in resumes and job offers.

Further, we also wanted to understand the incremental benefit of using simple rule-based systems, transitioning to semantic, and finally, supervised models, while keeping the explainability of the skill extraction outcomes under control. In Figure 1, we present an overview of the end-to-end system.

## 2.2 Taxonomies

We selected two taxonomies for the rule-based and semantic system, which model industrial and publicly available data to search for skills in text:

**ESCO_DB:** a collection of 3,039 occupations and 13,939 skills, translated in 28 official languages to standardize the European labor market.[5] This taxonomy is publicly available and widely used in previous work (Zhang et al., 2023; Decorte et al., 2023; Li et al., 2023). For our experiments, we extracted 131,623 entities relevant to the skill extraction task in English, with similar proportions across all languages.

**IN-HOUSE_DB:** a collection of 10,379 skill entities from our industrial partner. These entities were manually extracted from resumes and job offers from different European job markets. Each concept was also curated by specialized annotators and manually translated into more than 10 languages. For our study, we selected a subset of these languages due to the significant annotation effort required for producing reliable gold-standard data. Furthermore, we were also constrained by the available annotators for each language, the scope, and the timeline of the industrial project.

[5] https://esco.ec.europa.eu/en/use-esco/download

## 2.3 Datasets

For training our supervised models in Section 2.5, we selected job offers and resume datasets for both academic[6] and industrial settings. In particular, for our industrial dataset, **Arca24_JOB** and **Arca24_CV**, we obtained job offers and resumes from our industrial partner, who provides HR services and software to companies within the European market. From the provided data, we randomly sampled 100 resumes and 100 job offers for each of the 6 selected languages, resulting in a total of 1200 documents from ∼45 different domains (e.g., Engineering, Administration, Management). These documents were manually annotated by HR researchers and specialists with the following entities: *Degrees / Certifications*, *Domain*, *Hard Skills*, *Knowledge*, *Language*, *Occupations*, and *Soft Skills*. Further details of the annotation process are shared in Section 2.4.

All these datasets have different labeling schemas for each entity. Thus, we have remapped or discarded the original labels, considering only hard skills and occupations in this study. We show our datasets' entity distribution between hard skills and occupations, including total and unique counts and ratio between these two metrics and the average length of entities (by characters) in Table 1.

## 2.4 Annotation Guidelines

To identify skills within our dataset, we first developed well-defined annotation guidelines, relying on both in-house expertise (i.e., two HR researchers) and HRM literature, focusing on principles of KSAOs (Campion et al., 2011), which refers to knowledge, skills, abilities, and other at-

[6] We include the details of our selected public dataset **Green** in the Appendix A.2.1.

| Evaluation | | | | Exact (F1-score) | | | Partial (F1-score) | | | Time |
|---|---|---|---|---|---|---|---|---|---|---|
| DB | Test | Model | Lang. | Skills | Occ. | Overall | Skills | Occ. | Overall | (min) |
| IN-HOUSE | Green | Rule-based | EN | 0.145 | 0.275 | 0.179 | 0.258 | 0.419 | 0.300 | 5.200 |
| ESCO | | | | 0.166 | 0.355 | 0.216 | 0.361 | 0.493 | 0.396 | 57.800 |
| IN-HOUSE | | Semantic | | 0.135 | 0.280 | 0.173 | 0.265 | 0.483 | 0.322 | 5.600 |
| ESCO | | | | 0.172 | 0.210 | 0.185 | 0.369 | 0.333 | 0.357 | 25.600 |
| - | | mBERT | | 0.351 | 0.603 | **0.418** | 0.555 | 0.705 | **0.595** | 2.800 |

Table 2: Comparison of skill extraction methods. We present the F1-score (exact, partial) results on **Green** dataset (Job Offers) for Rule-Based/Semantic (taxonomy matching) vs. Supervised (labeled data fine-tuning) systems.

tributes. Each category was defined and illustrated with examples. The annotation team met in person to discuss and align on the definitions of skills in English and to test the guidelines[7] on a set of sample resumes. These initial annotations were collectively reviewed, and the guidelines were amended to clarify any ambiguous cases. Annotators then carried out a first round of annotations in their respective target languages, followed by an adjudication phase to resolve conflicts and ensure full alignment among the team.[8]

The labeling of entities was done by 4 annotators with previous experience in skills identification and/or annotation. All annotations were done using Docanno (Nakayama et al., 2018). The work of the annotators was distributed according to their language proficiency as follows: English-Spanish, French-German, Italian-Portuguese. Annotations in Italian, French, and Spanish were conducted by native speakers, whereas the others were done by annotators with a proficient C2 language level. Annotators gathered to review some uncertainties once all annotations had been conducted. Some degree of variability in annotations is likely to occur, as it reflects 1) variability in the content of the corpus of resumes and job offers annotated and 2) variability in the cultural norms and expectations of resumes and job offers across the six target languages.

## 2.5 Models

We experimented with various models that differ not only in complexity and level of explainability but also in resource requirements. While GPUs are commonly used, their implementation may not be feasible for all businesses due to their high resource consumption and associated cost. Moreover, our NLP application requires that every decision is fully traceable and explainable for system improvement. Relying on the nondeterministic output of conversational LLMs compromises this necessary level of auditability and explainability. Therefore, we selected a rule-based system, a semantic model, and a supervised system for our benchmark.[9] In the following section, we elaborate in the details of our supervised model and explain further on how it contributes better into the explainability of our results.

### 2.5.1 Supervised model

We selected multilingual BERT-base[10] as our main experimental model for our human-annotated data.[11] This model can both understand the languages under study and also use minimal GPU resources. Further, we chose this general-purpose model over a pre-trained skill extraction model given that the latter does not always perform better on the skill extraction task due to the learned biases from previous approaches (Vásquez-Rodríguez et al., 2024). Larger models could potentially yield better scores in a single domain, but the objective of this work is to assess the tangible value and customer benefit delivered by the incremental rule-based to supervised journey. For this, we use models that are resource-efficient enough (e.g., deployable locally without GPUs), rather than achieving state-of-the-art benchmarks. The deterministic nature of BERT-based models allows for controlled outputs, ensuring that decisions can be traced and

[7] Due to privacy and intellectual property concerns, we limit the disclosure of our guidelines. However, we describe our annotation strategies as thoroughly as possible to maximize reproducibility.

[8] Given the variety of the languages involved in the project and the proficiency level in various languages of the annotation team (consisting of 4 people), inter-agreement was not feasible for any other language than English. The focus was rather put on a very thorough training procedure and a strong alignment regarding annotations across the team.

[9] Further details of the rule-based and semantic systems are explained in the Appendix A.3.2.

[10] https://huggingface.co/google-bert/bert-base-multilingual-uncased

[11] The inclusion of multiple supervised baselines is planned as future work by our industrial partner. We prioritize demonstrating resource-efficient, incremental value over exhaustive comparative analysis. Moreover, the detailed investigation of supervised model variations has been the subject of our previously published work (Vásquez-Rodríguez et al., 2024).

explained to end users. In order to support the explainability of our approach, we have mapped predicted skills to an existing taxonomy, which indirectly justifies the reasoning behind each selection. Any future additions of skills to the model will be managed through a controlled process involving in-house annotation of previously undetected skills.

## 3 Experimental Setup

In this section, we present the experimental steps of our proposed methods, including the technical implementation details. For the rule-based and semantic systems, we have standardized the code into Python libraries to systematically run on the provided inputs. As for the supervised system, we performed fine-tuning of the model with the training and the validation set. We trained the skill extraction task as a span-based approach similar to Named Entity Recognition (NER)[12] where the final model was selected based on the F1-score on the validation set. We chose the best model for all the epochs run.

| | | Categories | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Lang.** | **Dataset** | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| EN (Green) | Job Offers | 45.1% | 4.5% | 3.6% | 6.3% | 15.3% | 9.0% | 16.2% |
| EN | Job Offers (our annotated data) | 15.8% | 5.2% | 9.5% | 3.1 % | 16.9% | 8.4% | 41.1% |
| FR | | 5.7% | 63.9% | 0.8% | 1.6% | 6.6% | 18.0% | 3.3% |
| IT | | 9.7% | 6.5% | 3.2% | 17.7% | 21.0% | 11.3% | 30.7% |
| DE | | 34.6% | 16.4% | 10.6% | 1.9% | 3.9% | 6.7% | 26.0% |
| PT | | 23.3% | 15.0% | 6.7% | 8.3% | 8.3% | 18.3% | 20.0% |
| ES | | 23.3% | 7.8% | 28.9% | 2.2% | 6.7% | 16.7% | 14.4% |
| EN | Resumes (our annotated data) | 26.6% | 5.3% | 10.6% | 1.1% | 3.2% | 17.0% | 36.2% |
| FR | | 39.3% | 0.9% | 29.5% | 0.0% | 4.5% | 8.9% | 17.0% |
| IT | | 37.9% | 6.0% | 4.3% | 0.9% | 3.45% | 24.1% | 23.3% |
| DE | | 31.8% | 15.1% | 13.5% | 1.6% | 6.4% | 9.5% | 22.2% |
| PT | | 44.0% | 2.4% | 0.0% | 2.4% | 0.0% | 26.4% | 24.8% |
| ES | | 18.4% | 1.0% | 12.2% | 1.0% | 7.1% | 12.2% | 48.0% |

Table 3: Results for the human evaluation

For the evaluation, we identified the entities in the corpora and evaluated the output IOB files against the reference using the *nervaluate* Python library.[13,14] As for the evaluation schema, we report both the exact and partial scores (i.e., the identification of at least one word of the target skill) to adjust to the challenge of finding fixed entities in such a variable task. Finally, we performed a human evaluation of the supervised system output in Section 3.1.[15]

### 3.1 Human Evaluation

We conducted a human evaluation on 30 sample sentences per language, per document type (i.e., job offers and resumes), for a total of 360 sentences (see Table 3). These samples were randomly drawn from the test set of supervised model outputs. The primary goal was to analyze error patterns and identify possible drawbacks of neural models. To focus on the error analysis, we filtered out all correctly predicted samples. The remaining predictions and their corresponding references were displayed side by side in IOB format, token by token.

The human evaluation was carried out by three evaluators with prior experience in skill annotation. When possible, we would assign different annotators for labeling and human evaluation. This task assessed possible errors between the predicted output and the reference. Each entity is classified into different error categories (Nguyen et al., 2024), which are further explained in Section 3.1.1.

#### 3.1.1 Human Evaluation Categories

For the human evaluation task, we categorized the differences between the predicted output and the reference. Each entity could be classified into the following categories: 1) *Skill definition misalignment*, when the prediction includes a career-related term that is not a skill; 2) *Wrong extraction*, when the detected entity is entirely unrelated to any skill; 3) *Conjoined skills*, when two distinct skills are incorrectly merged into one (e.g., develop report software and statistical software should be treated as two separate skills); 4) *Extended span*, when the predicted entity is longer than the correct reference; 5) *Incorrect annotations*, when the human annotation itself is imprecise, containing incorrect values in the reference; 6) *Other* cases, such as when the prediction is shorter than the reference or when the reference skill was not predicted at all. Additionally, we also label the correct entities when the prediction is equal to the reference. We refer to this classification as category 0.

## 4 Results and Discussion

We present our results evaluating the performance of the skills detection task using F1-score with an

[12] We used a NER-based evaluation to leverage existing resources and tools. Despite challenges in consistently defining concept boundaries, our results align with prior work.

[13] https://pypi.org/project/nervaluate/

[14] We selected IOB standard due to the availability of corpora for this task and for its compatibility with the annotations output format.

[15] We report our data processing steps and training parameters in the Appendix A.

| Evaluation | | | | Exact (F1-score) | | | Partial (F1-score) | | | Time |
|---|---|---|---|---|---|---|---|---|---|---|
| DB | Test | Model | Lang. | Skills | Occ. | Overall | Skills | Occ. | Overall | (min) |
| IN-HOUSE | Jobs (our annotated data) | Rule-based | EN | 0.043 | 0.720 | 0.108 | 0.170 | 0.840 | 0.235 | 0.535 |
| ESCO | | Rule-based | | 0.047 | 0.270 | 0.078 | 0.230 | 0.297 | 0.239 | 6.000 |
| IN-HOUSE | | Semantic | | 0.016 | 0.667 | 0.075 | 0.165 | 0.792 | 0.221 | 2.300 |
| ESCO | | Semantic | | 0.014 | 0.142 | 0.038 | 0.189 | 0.157 | 0.183 | 17.900 |
| - | | mBERT | | 0.220 | 0.560 | **0.247** | 0.405 | 0.680 | **0.427** | 1.900 |
| IN-HOUSE | | Rule-based | FR | 0.000 | 0.000 | 0.000 | 0.065 | 0.172 | 0.081 | 0.610 |
| ESCO | | Rule-based | | 0.022 | 0.037 | 0.024 | 0.160 | 0.222 | 0.170 | 2.600 |
| IN-HOUSE | | Semantic | | 0.018 | 0.118 | 0.031 | 0.156 | 0.353 | 0.182 | 3.000 |
| ESCO | | Semantic | | 0.045 | 0.071 | 0.050 | 0.202 | 0.235 | 0.207 | 7.600 |
| - | | mBERT | | 0.443 | 0.591 | **0.464** | 0.595 | 0.659 | **0.605** | 2.200 |
| IN-HOUSE | | Rule-based | IT | 0.022 | 0.296 | 0.083 | 0.172 | 0.481 | 0.242 | 0.370 |
| ESCO | | Rule-based | | 0.017 | 0.065 | 0.027 | 0.214 | 0.226 | 0.216 | 1.300 |
| IN-HOUSE | | Semantic | | 0.038 | 0.345 | 0.086 | 0.274 | 0.552 | 0.317 | 2.600 |
| ESCO | | Semantic | | 0.034 | 0.067 | 0.045 | 0.267 | 0.157 | 0.230 | 7.100 |
| - | | mBERT | | 0.135 | 0.615 | **0.235** | 0.419 | 0.692 | **0.476** | 2.800 |
| IN-HOUSE | | Rule-based | DE | 0.085 | 0.296 | 0.124 | 0.178 | 0.444 | 0.228 | 0.357 |
| ESCO | | Rule-based | | 0.036 | 0.303 | 0.097 | 0.098 | 0.333 | 0.152 | 0.768 |
| IN-HOUSE | | Semantic | | 0.106 | 0.529 | 0.166 | 0.213 | 0.706 | 0.282 | 2.700 |
| ESCO | | Semantic | | 0.069 | 0.186 | 0.104 | 0.202 | 0.244 | 0.215 | 6.900 |
| - | | mBERT | | 0.299 | 0.667 | **0.367** | 0.476 | 0.738 | **0.524** | 2.200 |
| IN-HOUSE | | Rule-based | ES | 0.022 | 0.000 | 0.021 | 0.110 | 0.200 | 0.115 | 0.352 |
| ESCO | | Rule-based | | 0.027 | 0.000 | 0.025 | 0.195 | 0.231 | 0.198 | 1.800 |
| IN-HOUSE | | Semantic | | 0.039 | 0.000 | 0.037 | 0.221 | 0.333 | 0.225 | 2.200 |
| ESCO | | Semantic | | 0.042 | 0.000 | 0.036 | 0.211 | 0.108 | 0.197 | 9.400 |
| - | | mBERT | | 0.355 | 0.800 | **0.378** | 0.505 | 0.800 | **0.520** | 1.800 |
| IN-HOUSE | | Rule-based | PT | 0.028 | 0.095 | 0.037 | 0.099 | 0.143 | 0.104 | 0.387 |
| ESCO | | Rule-based | | 0.068 | 0.258 | 0.092 | 0.179 | 0.323 | 0.197 | 1.400 |
| IN-HOUSE | | Semantic | | 0.116 | 0.483 | 0.155 | 0.252 | 0.586 | 0.288 | 2.500 |
| ESCO | | Semantic | | 0.065 | 0.167 | 0.076 | 0.173 | 0.278 | 0.185 | 6.100 |
| - | | mBERT | | 0.444 | 0.357 | **0.429** | 0.595 | 0.500 | **0.578** | 2.100 |

Table 4: Comparison of skill extraction methods. We present the F1-score (exact, partial) results on the **Job Offers** dataset for Rule-Based/Semantic (taxonomy matching) vs. Supervised (mBERT, labeled data fine-tuning) systems.

exact match assessment. In Table 4, we show supervised systems trained in job offers that achieve the highest score in French, followed by Portuguese and English-Green. For the state-of-the-art results, we present the Green dataset in Table 2, where, for English, results remained competitive throughout. For resumes, supervised systems also perform well, as observed in Portuguese, followed by Italian, and English in Table 5. These results aim to characterize the breadth of variability real-word skill descriptions across document types (job offers vs resumes), domains, and different languages, as opposed to optimizing for benchmark performance on the skill extraction task. And consequently, these show how approaches that are currently being used in industry fare on these data types. Furthermore, we can confirm that the state-of-the-art performance on the academic Green Dataset (Nguyen et al., 2024) is comparable to our supervised model, which is expected to be higher ($\sim$10%) compared to our F-measure due to the choice of a smaller multilingual model to support a minimal resource setting with multiple languages.

In relation to the detection of occupations, results are more straightforward, yielding scores of up to 0.8 that are significantly higher than those of hard skills, when considering the best performance in the partial evaluation. Most methods demonstrate efficient memory consumption, with CPU-based approaches requiring a maximum of $\sim$5 GB and GPU-based methods utilizing up to $\sim$12.85 GB of RAM. Similarly, runtime remain reasonable across all approaches, except for the rule-based method on the Green dataset, which can take up to an hour due to its size. In industrial settings, optimizing resources and performance is crucial. Poor real-time skill extraction performance can negatively impact user experience and the management of large-scale data. Additionally, scaling up resources (e.g., adding more RAM) can significantly increase costs. Therefore, the trade-offs between different approaches, such as CPU-based vs. GPU-based methods, must be carefully justified.

Our study highlights key differences between job offers and resumes (Section 4.1), annotation difficulties (Section 4.2), multilingual considerations (Section 4.3), the trade-off between explainability, generalization and fairness (Section 4.4) and challenges involved in the transition of research into NLP applications (Appendix A.1).

| Evaluation | | | | Exact (F1-score) | | | Partial (F1-score) | | | Time |
|---|---|---|---|---|---|---|---|---|---|---|
| DB | Test | Model | Lang. | Skills | Occ. | Overall | Skills | Occ. | Overall | (min) |
| IN-HOUSE | Resumes (our annotated data) | Rule-based | EN | 0.160 | 0.226 | 0.171 | 0.253 | 0.481 | 0.289 | 2.000 |
| ESCO | | Rule-based | | 0.159 | 0.100 | 0.144 | 0.267 | 0.217 | 0.254 | 17.500 |
| IN-HOUSE | | Semantic | | 0.130 | 0.236 | 0.146 | 0.217 | 0.535 | 0.263 | 3.900 |
| ESCO | | Semantic | | 0.119 | 0.074 | 0.106 | 0.217 | 0.156 | 0.200 | 19.700 |
| - | | mBERT | | 0.382 | 0.400 | **0.385** | 0.487 | 0.531 | **0.495** | 2.100 |
| IN-HOUSE | | Rule-based | FR | 0.026 | 0.260 | 0.068 | 0.124 | 0.410 | 0.175 | 2.600 |
| ESCO | | Rule-based | | 0.050 | 0.107 | 0.062 | 0.191 | 0.321 | 0.218 | 10.300 |
| IN-HOUSE | | Semantic | | 0.057 | 0.281 | 0.093 | 0.184 | 0.545 | 0.242 | 3.600 |
| ESCO | | Semantic | | 0.044 | 0.050 | 0.046 | 0.190 | 0.249 | 0.203 | 8.000 |
| - | | mBERT | | 0.252 | 0.364 | **0.269** | 0.407 | 0.580 | **0.434** | 2.200 |
| IN-HOUSE | | Rule-based | IT | 0.111 | 0.442 | 0.197 | 0.258 | 0.565 | 0.338 | 0.917 |
| ESCO | | Rule-based | | 0.023 | 0.425 | 0.125 | 0.200 | 0.523 | 0.282 | 3.900 |
| IN-HOUSE | | Semantic | | 0.095 | 0.444 | 0.171 | 0.295 | 0.632 | 0.369 | 3.600 |
| ESCO | | Semantic | | 0.024 | 0.203 | 0.076 | 0.252 | 0.334 | 0.276 | 7.600 |
| - | | mBERT | | 0.405 | 0.444 | **0.414** | 0.597 | 0.606 | **0.599** | 3.300 |
| IN-HOUSE | | Rule-based | DE | 0.069 | 0.448 | 0.168 | 0.151 | 0.510 | 0.245 | 0.763 |
| ESCO | | Rule-based | | 0.090 | 0.122 | 0.100 | 0.194 | 0.223 | 0.204 | 1.500 |
| IN-HOUSE | | Semantic | | 0.124 | 0.613 | 0.226 | 0.228 | 0.717 | 0.329 | 3.700 |
| ESCO | | Semantic | | 0.074 | 0.165 | 0.107 | 0.197 | 0.231 | 0.210 | 7.800 |
| - | | mBERT | | 0.257 | 0.459 | **0.315** | 0.426 | 0.574 | **0.468** | 2.400 |
| IN-HOUSE | | Rule-based | ES | 0.045 | 0.333 | 0.085 | 0.132 | 0.476 | 0.179 | 1.200 |
| ESCO | | Rule-based | | 0.017 | 0.299 | 0.061 | 0.115 | 0.463 | 0.170 | 3.200 |
| IN-HOUSE | | Semantic | | 0.085 | 0.435 | 0.128 | 0.228 | 0.587 | 0.272 | 3.000 |
| ESCO | | Semantic | | 0.042 | 0.139 | 0.065 | 0.206 | 0.222 | 0.210 | 10.300 |
| - | | mBERT | | 0.265 | 0.509 | **0.301** | 0.382 | 0.582 | **0.411** | 4.200 |
| IN-HOUSE | | Rule-based | PT | 0.029 | 0.217 | 0.058 | 0.141 | 0.337 | 0.172 | 1.300 |
| ESCO | | Rule-based | | 0.012 | 0.155 | 0.038 | 0.127 | 0.250 | 0.150 | 3.800 |
| IN-HOUSE | | Semantic | | 0.038 | 0.279 | 0.070 | 0.185 | 0.589 | 0.239 | 4.100 |
| ESCO | | Semantic | | 0.022 | 0.114 | 0.040 | 0.195 | 0.247 | 0.205 | 7.900 |
| - | | mBERT | | 0.483 | 0.532 | **0.491** | 0.598 | 0.662 | **0.608** | 2.400 |

Table 5: Comparison of skill extraction methods. We present the F1-score (exact, partial) results on the **Resumes** dataset for Rule-Based/Semantic (taxonomy matching) vs. Supervised (mBERT, labeled data fine-tuning) systems.

### 4.1 Job offers vs Resumes

Job offers are more standardized than resumes, as organizations often follow similar guidelines for all their job openings. However, they often contain multiple career-related terms that are not necessarily related to skills specifically required for the position at hand, as shown in Table 3 (category 1), which can pose challenges for skill annotation and extraction. Job offers are also much more available for collection and analysis than resumes. This has two important consequences. First, their prevalence and availability on the internet could contribute to higher performance on publicly available datasets for English job offers, such as Green (Table 2). In contrast, performance is lower for IN-HOUSE job offers as shown in Table 4.

Second, academic research has focused on job offers for availability reasons. However, job offers are very distinct from resumes in how they are structured, in the way they formulate skills and experience, and in the information they contain. This is bound to eventually limit the quality of the systems if trained on a non-representative dataset. A notable difference is seen in the partial evaluation of the supervised method. The success of skill detection varies between resumes and job offers, with some cases achieving a balanced trade-off between precision and recall, while others do not.

Furthermore, job offers and resumes vary across industries and languages. Although we used random sampling to select a meaningful population of job offers and resumes, the presence of skills varies according to the domain, their relative occupation, and the target language. Some job advertisements are concise, spanning only a few lines, while others are more detailed. A similar pattern is observed in resumes, which are highly personal and vary in format and style. For example, certain domains such as technology-oriented datasets (i.e., Green) show that skills are more straightforward to identify (see Table 6). In contrast, domains such as management might need a higher level of inference. Moreover, language and nationality play a role as norms on what to put in a resume, and what is legally required to write in a job offer varies.

### 4.2 Annotations Challenges

Although we established clear annotation guidelines based on academic literature and validated these with HR experts, the variability of resumes

across disciplines (e.g., software developer vs manager) and languages still poses a challenge for skill annotation consistency. As an example, we present the experiments for Italian job offers (F1-score: 0.235), where annotation quality significantly contributed to the poor performance. A manual review revealed that entity annotations in Italian were long, containing a higher number of words compared to other languages, which severely impacted the effectiveness of the entity-oriented detection approach. Our error analysis highlights how inconsistencies in gold annotations (see Table 3, category 5) can also impact model performance. The detailed error analysis shows that resumes present greater challenges, with most errors falling into Category 5 (annotation errors) or Category 6 ("other"), where the model struggles not only to predict a skill but also to align with the reference annotations (e.g., shorter annotations). The definition of skill is complex, whereas occupations are more straightforward to identify in both resumes and job offers, as seen when comparing the F1-scores of "Skills" and "Occ." in Tables 2, 4 and 5. We additionally show examples of these challenges in Table 6 and 7 (Appendix), where precise skill boundaries are challenging to define.

| Model | Taxonomy | Example |
|---|---|---|
| Reference | - | From my experience, I learned and used different languages PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C |
| Rule-based | IN-HOUSE | From my experience, I learned and used different languages PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C |
| Rule-based | ESCO | From my experience, I learned and used different languages PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C |
| Semantic | IN-HOUSE | From my experience, I learned and used different languagues PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C |
| Semantic | ESCO | From my experience, I learned and used different languages PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C |
| mBERT | - | From my experience, I learned and used different languages PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C |

Table 6: System outputs from English resumes

### 4.3 Multilinguality

Rule-based and semantic systems demonstrated similar performance (measured by exact F1-score) across languages (up to ∼0.2), while more significant performance differences were observed in the supervised results for job offers in Table 4, comparing Italian and French. Overall, the mBERT multilingual model outperformed rule-based and semantic systems. From our error analysis, we can notice that 1) In certain industries, the labor market is standardized to English, leading to a mix of local language and English terms within job descriptions, which could have a positive or negative impact in performance depending on the selected multilingual system; 2) English resumes are likely to be written by non-native speakers, which potentially can influence the quality of the submitted resume.[16]

### 4.4 Explainability, Generalization, and Fairness

Rule-based and semantic methods show a lower performance compared to supervised approaches, but offer high explainability because extracted skills can be directly linked to a curated taxonomy and are relatively easy to adapt, albeit at the cost of labor-intensive and expensive taxonomy maintenance. Supervised systems demonstrate better generalization, but their performance is highly dependent on the quality, quantity, and diversity of the data, as evidenced by the discussed results.

Taxonomy-driven methods are inherently fair, as concepts remain consistent across target languages. In contrast, neural methods may introduce cultural biases, potentially favoring or disadvantaging specific candidates. In some cases, a detailed taxonomy, such as the ESCO taxonomy, can enhance performance by using a rule-based system and more standardized job offers. However, for standalone academic datasets, the results appear to be "artificially" high, as in Green for partial evaluation. The large difference between this dataset and our multilingual datasets from actual client data suggests that the skill extraction task may be harder than one would expect from academic data, which is often limited to a particular domain, language, and to the relevant phrases with explicit skill mentions. For practical purposes, it is not necessary to exhaustively detect every skill, as many are redundant. A representative sample is often sufficient to provide a clear sense of the candidate's profile and to support manual candidate analysis. Therefore, we conducted a human error analysis to analyze the quality of system outputs and the distribution of errors. As shown in Table 3, for English job offers (Green), nearly 50% of entities were correctly identified (Category 0), suggesting a lower level of complexity in these documents compared to resumes, which had a success rate of only 26.60%.

[16] We further discuss the linguistic and cultural aspects of the skill extraction task in the Limitations section.

We found no conclusive evidence linking specific error types to the multilingual aspect, particularly given the additional differences in domain distribution across languages. Except for French and Spanish, most failures in the supervised models fall into categories 5 and 6, where errors stem from incorrect human annotations, missing predictions, or shorter-than-expected extractions. The significance of these analyses extends beyond research, as fair and reliable candidate selection remains a highly sensitive task, requiring human oversight to ensure trustworthiness and mitigate biases in automated hiring systems. The controllability and explainability of these systems are essential for the engagement and trust of stakeholders.

### 4.5 Reproducibility

Due to the proprietary nature of the job offers and resumes provided by our industrial partner, the release of the primary dataset is restricted. We evaluated pseudonymization as an alternative; however, the high density of sensitive personal information within resumes makes full anonymization and subsequent public release complex from both legal and ethical perspectives.

To ensure reproducibility while respecting these constraints, we adopt the following strategies:

**1) Experiments on public data:** To mitigate privacy constraints, we additionally report results using public job offers and taxonomies (e.g., the Green dataset and ESCO taxonomy);

**2) Methodological transparency:** Our study encompasses hard-skill extraction across job offers and resumes, utilizing multiple languages and resource-wise incremental baselines. We also include comprehensive details of our methodology to ensure these benchmarks remain as reproducible as possible, providing a transparent framework for our real-world skill extraction approach.

While we acknowledge the trade-offs inherent in using proprietary data, this study offers a rare, large-scale analysis of skill extraction in authentic industrial scenarios. We believe the broad scope and practical insights of this research provide significant value to the community beyond what is achievable with restricted or synthetic datasets.

## 5 Conclusion

Traditionally, the skill extraction task has been approached with a limited scope, focusing on one or two languages, exclusively analyzing job offers, or relying on resource-intensive methods. Our work presents the first experiment of its kind that proposes and integrates different methodologies at multiple levels of complexity, languages, and side-by-side job offer and resume data through an extensive annotation effort covering 1200 resumes and job offers. We achieved an improved performance over production across different system variations, evaluation schemes, and data settings, detected a higher volume of qualitative instances, and provided valuable insights to guide future advancements. Finally, we present our discussions on future work in the Appendix A.4.

### Limitations

In this section, we outline the limitations of our research in multiple aspects to be considered as follows:

**Data release:** As discussed in Section 4.5, we present our results using publicly available job offers as well, as our analysis builds on proprietary job offers and resumes from our industrial partner. However, we provide a detailed description of our research methodology to enhance reproducibility.

**Taxonomies distribution:** Our taxonomies are fundamentally different and non-standardized between them. The ESCO DB was created to standardize skills for the European job market, but it presents a formalized approach that does not reflect how people express skills in real-world resumes. In contrast, the IN-HOUSE taxonomy is crafted directly from the actual occurrences within real resumes and job offers, making it more practical. The ESCO DB is freely available. However, we refrain from providing specific distribution details of the IN-HOUSE taxonomy due to proprietary and privacy concerns.

**NER evaluation:** Another limitation is the use of the NER approach for evaluation, which allows the use of existing resources to assess the quality of our annotations and the use of open-source tools for entity labeling. However, we acknowledge that it is difficult to capture the boundaries of each concept given the variable and subjective nature of the labeling task, resulting in overall lower scores, but still consistent with previous work.

**Cultural nuances in the annotation task:** We acknowledge that variations in hiring cultures across domains and languages can impact our work.

However, a comprehensive evaluation of these nuances would require a dedicated study that falls outside the scope of the current paper.

**Annotation-related performance differences:** We follow up on the discussion of the differences in performance in language as Italian, which are fundamentally attributed to the variability in annotation. We established rigorous, consistent guidelines and performed our annotations with in-house domain specialists instead of crowdworkers. Still, it cannot eliminate the inherent subjectivity present in human annotation. We have included this variability in our experiments to reflect real-world complexity. The persistence of this challenge for many years underscores why skill extraction remains a difficult research task.

### Ethical Considerations

We ensure strict adherence to the corresponding data and privacy through non-disclosure agreements. Furthermore, we have obtained all necessary licenses and data consents for managing job offers and resume information in compliance with relevant regulations. Additionally, no data was published in the cloud; all operations were handled locally on premises. Furthermore, the final output of the skill extraction systems was always mapped to a controlled, real-world taxonomy, which effectively prevented the exposure of any sensitive data derived from the raw source. However, due to the sensitive nature of these resources and the ethical considerations involved, we are unable to share them publicly. Furthermore, we recognize the importance of maintaining fairness, transparency, and accountability in automated hiring systems. Our approach is designed to eventually support recruiters by providing them with additional insights, without delegating candidate selection decisions to any algorithm. This ensures that final hiring decisions remain under human supervision, mitigating potential biases and ethical concerns.

## References


David G. Allen and James M. Vardaman. 2017. Recruitment and retention across cultures. *Annual Review of Organizational Psychology and Organizational Behavior*, 4(1):153–181.

David Beauchemin, Julien Laumonier, Yvan Ster, and Marouane Yassine. 2022. "fijo": a french insurance soft skill detection dataset. *arXiv*.

Akshay Bhola, Kishaloy Halder, Animesh Prasad, and Min-Yen Kan. 2020. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5832–5842, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. 2018. Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, 86:319–328.

Michael A. Campion, Alexis A. Fink, Brian J. Ruggeberg, Linda Carr, Guy M. Phillips, and Ronald B. Odman. 2011. Doing competencies well: Best practices in competency modeling. *Personnel Psychology*, 64(1):225–262.

Benjamin Clavié and Guillaume Soulié. 2023. Large language models as batteries-included zero-shot esco skills matchers. In *RecSys in HR'23: The 3rd Workshop on Recommender Systems for Human Resources, in conjunction with the 17th ACM Conference on Recommender Systems, September 18–22, 2023, Singapore, Singapore.*

Jens-Joris Decorte, Severine Verlinden, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Extreme multi-label skill extraction training using large language models. In *International Workshop on AI for Human Resources and Public Employment Services (AI4HR&PES) at ECML-PKDD*.

Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2025. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Trans. Intell. Syst. Technol.*, 16(1).

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of llm agents in recruitment: A novel framework for resume screening. *Preprint*, arXiv:2401.08315.

Ann-sophie Gnehm, Eva Bühlmann, Helen Buchs, and Simon Clematide. 2022. Fine-grained extraction and classification of skill requirements in German-speaking job ads. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 14–24, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Green, Diana Maynard, and Chenghua Lin. 2022. Development of a benchmark corpus to support entity recognition in job descriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1201–1208, Marseille, France. European Language Resources Association.

Akshay Gugnani and Hemant Misra. 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.

Shiqiang Guo, Folami Alamudun, and Tracy Hammond. 2016. Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60:169–182.

Faizan Javed, Phuong Hoang, Thomas Mahoney, and Matt McNair. 2017. Large-scale occupational skills normalization for online recruitment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(2):4627–4634.

Ilkka Kivimäki, Alexander Panchenko, Adrien Dessy, Dries Verdegem, Pascal Francq, Hugues Bersini, and Marco Saerens. 2013. A graph-based approach to skill extraction from text. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, pages 79–87, Seattle, Washington, USA. Association for Computational Linguistics.

Sunil Kumar Kopparapu. 2010. Automatic extraction of usable information from unstructured resumes to aid search. In *2010 IEEE International Conference on Progress in Informatics and Computing*, volume 1, pages 99–103.

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Nan Li, Bo Kang, and Tijl De Bie. 2023. Skillgpt: a restful api service for skill extraction and standardization using a large language model. In *International Workshop on AI for Human Resources and Public Employment Services (AI4HR&PES) at ECML-PKDD*.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Khanh Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. Rethinking skill extraction in the job market domain using large language models. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 27–42, St. Julian's, Malta. Association for Computational Linguistics.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 1–15, St. Julian's, Malta. Association for Computational Linguistics.

Damian A. Tamburri, Willem-Jan Van Den Heuvel, and Martin Garriga. 2020. Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 391–394.

Laura Vásquez-Rodríguez, Bertrand Audrin, Samuel Michel, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa, and Lonneke Van Der Plas. 2024. Hardware-effective approaches for skill extraction in job offers and resumes. In *RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems, October 14–18, 2024, Bari, Italy.* CEUR Workshop Proceedings.

Laura Vásquez-Rodríguez, Bertrand Audrin, Samuel Michel, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa, and Lonneke van der Plas. 2024. Hardware-effective approaches for skill extraction in job offers and resumes. In *Proceedings of the 4th Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2024) co-located with the 18th ACM Conference on Recommender Systems (RecSys 2024)*, volume 3788 of *CEUR Workshop Proceedings*, pages 1–12, Bari, Italy. CEUR-WS.org.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. 2022a. SkillSpan: Hard and soft skill extraction from English job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4962–4984, Seattle, United States. Association for Computational Linguistics.

Mike Zhang, Kristian Nørgaard Jensen, and Barbara Plank. 2022b. Kompetencer: Fine-grained skill classification in Danish job postings via distant supervision and transfer learning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 436–447, Marseille, France. European Language Resources Association.

Mike Zhang, Kristian Nørgaard Jensen, Rob van der Goot, and Barbara Plank. 2022c. Skill extraction

from job postings using weak supervision. In *RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA*. CEUR Workshop Proceedings.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2023. ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11871–11890, Toronto, Canada. Association for Computational Linguistics.

Meng Zhao, Faizan Javed, Ferosh Jacob, and Matt McNair. 2015. Skill: A system for skill identification and normalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(2):4012–4017.

## Acknowledgments

We would like to thank Arca24 HR specialists for their support with annotations in this project. Finally, we gratefully acknowledge the support from Innosuisse (grant 104.069 IP-ICT).

# A Appendix

## A.1 Research on NLP applications

We reflect on the challenges of translating NLP research into industry applications. In our study, we carried out experiments within the framework of an industrial project aimed at developing improved algorithms for candidate selection. In this context, we faced a clear trade-off between the complexity of the NLP methods and the state of the art of research and the pragmatic requirements of the industry regarding a deployable solution within existing systems over a limited timeframe. The project has now been completed, and our industrial partner has successfully integrated the algorithms. This outcome demonstrates not only the practical readiness of the research but also underlines a key lesson: the importance of a bottom-up development process, where all industry parameters need to be accounted for beyond raw performance taking into account the available infrastructure, costs and time constraints, and explainability for the users needs, costs, and performance are defined collaboratively to align with the company's roadmap, while at the same time enabling innovation through research that is explainable and fair.

## A.2 Methodology

In this section, we discuss the freely available data related the Green dataset (Section A.2.1). Finally, we present our data preprocessing steps (Section A.3.2), the description of our rule-based and semantic models (Section A.2.1), including the training details for the supervised system (Section A.3.3).

### A.2.1 Publicly available data

We present the selected academic dataset (Green) for our experiments:

**Green_JOB (Green et al., 2022):** This academic dataset is a set of job offers from UK job boards with 8670 sentences for training, 964 for validation, and 335 for testing.[17] The annotations of these jobs were done by crowdsourcing efforts, labeling entities with the following types: *Skills*, *Knowledge*, *Occupation*, *Experience*, and *Domain*. We selected this dataset as a reference to previous work.

## A.3 Experimental Setup

In this section, we extend the implementation details of the paper, including the data processing steps (Section A.3.1), the training parameters (Section A.3.3), and the selected categories for the human evaluation (Section 3.1.1).

### A.3.1 Data Preprocessing

We preprocessed the taxonomy to support the rule-based system by precalculating all terms and storing them in a local JSON file. This allows for efficient matching of all possible term variants for a given word (e.g., Software developers → software developers). For the semantic system, we did not change the original terms as the semantic similarity will generalize enough to capture these minor syntactic differences.

Concerning the datasets, we divided the texts presented in Section 2.3, for both job offers and resumes, into 80% for training, 10% for validation, and 10% for testing. In particular, for the supervised system, we split documents into 200 tokens to avoid any issues with the input size of the multilingual models. For the Green dataset, we kept the original distribution of the splits. All systems were evaluated on the same test set and in 6 languages. As for the input and output format for all systems, we relied on the IOB schema (Ramshaw and Marcus, 1999) to clearly understand the boundaries of each entity.

### A.3.2 Models

In this section, we explain the implementation details of the rule-based and supervised system.

[17] https://huggingface.co/datasets/jjzha/green

**Rule-based:** We followed the implementation of the rule-based system proposed by Vásquez-Rodríguez et al. (2024) with minor enhancements. The rule-based system will search each term from the taxonomies (see Section 2.2) in the text. To maximize the matching of concepts, the taxonomy is normalized using techniques such as lemmatization and stemming. The final list of matching concepts is ranked according to their similarity using semantic similarity and Levenshtein distance (Levenshtein, 1966), as overlapping of n-grams can happen.

**Semantic SBERT:** We present the semantic system as a better generalization of the rule-based algorithm. To achieve this, we have improved the system proposed by Vásquez-Rodríguez et al. (2024), replacing the similarity mechanism from Spacy models[18] for SBERT embeddings (Reimers and Gurevych, 2020). Sentence embedding comparisons are more accurate and showed to be 36X faster than spacy-based methods based on the evaluation of the *Green_JOB* test set. In comparison to the rule-based systems, the semantic model can find concepts that are closely related rather than an exact match. The main benefit is that there is no need for an exhaustive taxonomy, which in that case, degrades the performance of the algorithm as it grows. Finally, for the semantic comparisons, we selected the *distiluse-base-multilingual-cased-v1* model,[19] a multilingual model supporting 15 languages, distilled from the universal multilingual sentence encoder proposed by Yang et al. (2020).

### A.3.3 Training Parameters

We conducted experiments using three random seeds and reported the average results across all runs. The selected hyperparameters include a batch size of 16, a learning rate of $5.0 \times 10^{-5}$, and a maximum of 10 epochs. For training, jobs were executed on heterogeneous hardware configurations, using either four Intel(R) Xeon(R) Platinum 8468 or AMD EPYC 7742 CPU cores, 16 GB of RAM, and one of two types of GPUs: an NVIDIA V100 with 32 GB of memory or an NVIDIA RTX 3090 with 24 GB of memory. Evaluation was carried out on a separate setup with two AMD EPYC 7742 CPU cores and 32 GB of RAM.

[18] https://spacy.io/models/

[19] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

| Model | Taxonomy | Example |
|---|---|---|
| Reference | - | An analytical approach to problem solving excellent team working skills |
| Rule-based | IN-HOUSE | An analytical approach to problem solving excellent team working skills |
| Rule-based | ESCO | An analytical approach to problem solving excellent team working skills |
| Semantic | IN-HOUSE | An analytical approach to problem solving excellent team working skills |
| Semantic | ESCO | An analytical approach to problem solving excellent team working skills |
| mBERT | - | An analytical approach to problem-solving excellent team working skills |

Table 7: System outputs from the English job offers.

## A.4 Future Work

We discuss our future work in terms of the inclusion skills that arise in the job market, the cultural nuances that can affect our work, and possible approaches to avoid redundancy between taxonomies.

**Inclusion of new skills:** The inclusion of novel, in-domain skills is done by continuously analyzing skills extracted by the supervised systems that are currently missing from the core taxonomy. This approach offers an improved selection method over the traditional, costly, fully manual method performed by specialists and accelerates the process by already suggesting potential skills that later will be curated by humans. Only when a completely new or unfamiliar domain arises, specific re-annotation efforts are necessary to maintain accuracy and performance. The unsupervised detection of new skills and the integration of large LLMs for this purpose remain outside the current scope of our project.

**Cultural nuances:** Beyond this work, there remain interesting avenues for exploration, particularly concerning the cultural and linguistic aspects of skill expression (Allen and Vardaman, 2017). The differences in recruiting culture significantly influence resume structure and how strengths are expressed. Investigating these cultural nuances is a promising direction for future research.

**Taxonomies alignment:** Given the public access to both taxonomies, we could compute semantic similarity between entries to identify redundant occurrences of existing skills. In addition, it would be very useful to paraphrase and standardize the concepts into a more realistic, resume-like style using LLMs, assuming that sufficient resources are available, that document privacy can be preserved, and hallucinations are mitigated.

Figure 2: Example of the annotation effort using the Docanno tool. Data was anonymized to avoid correlations to a real candidate.