Multilingual Skill Extraction

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

Innovation project supported by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Innosuisse – Swiss Innovation Agency

Skill extraction methods in **academic research** are frequently limited to **English job offers**, failing to capture the complexity of global recruitment. This study directly addresses this limitation. We evaluate multiple skill extraction methods across **six languages**, comparing **resumes** and **job offers** across diverse domains to build a more comprehensive, industry-relevant **framework**.

Summary

- **Data:** 1,200 total documents across 6 languages: English, French, German, Italian, Spanish, and Portuguese.
- **Split:** 100 job offers and 100 resumes per language in approximately 45 different professional domains.
- **Human-in-the-Loop:** Resumes and job offers were manually annotated by HR specialists using the Docanno tool.
- **HR Framework:** Skills were mapped using HRM literature and in-house expertise, utilizing the KSAO framework (Knowledge, Skills, Abilities, and Other attributes; Campion et al., 2011).

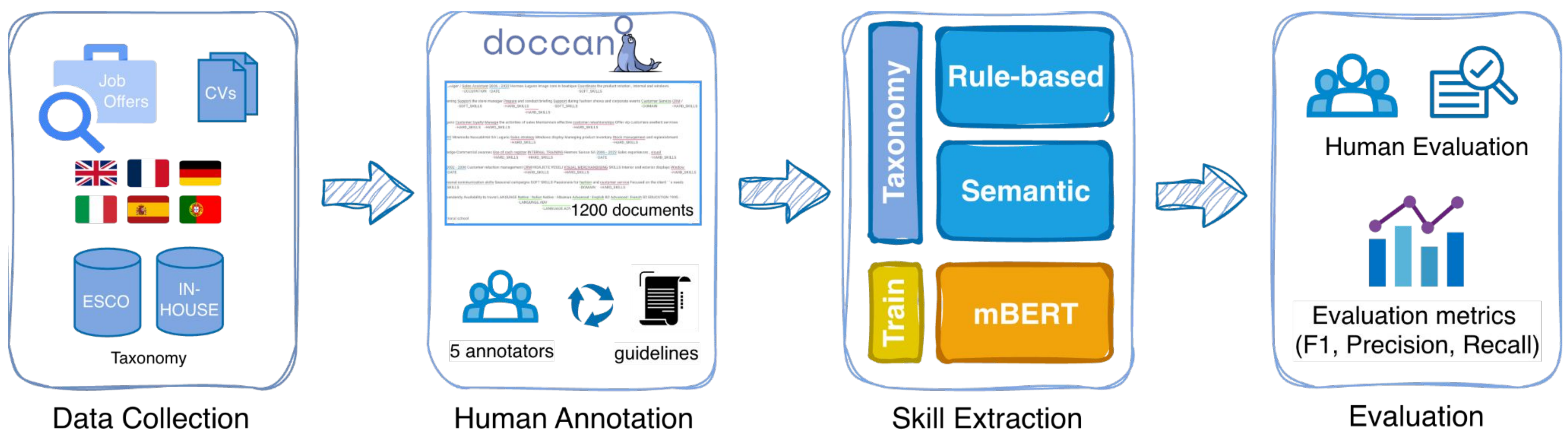

Figure 1. End-to-End Skill Extraction System

Multilingual Skill Extraction

Method	How?	Pros (+)	Cons (-)
Rule-based system	**Strict matching** using in-house (industry) and ESCO taxonomy	Fully **transparent**, taxonomy that is easy to update manually	Poor **variability** handling
Semantic (sentence embeddings)	Sentence embeddings to **find** closely **related terms**	Better **generalization**, extracted skills still traceable via the taxonomy	Can**not** identify completely **new concepts**
Supervised mBERT	Fine-tuned, **multilingual** BERT-base model **trained** with manual **annotations**	Captures **context** across languages	**"Black-box"** model; harder to manipulate

Table 1. Skill Extraction Methods Comparison

Skill Extraction Examples

Job Offers (EN) - **Semantic (in-house)**

An analytical approach to problem solving excellent team working skills
An analytical approach to problem solving excellent team working skills

Resumes (EN) - **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
From my experience, I learned and used different languages PHP HTML CSS Javascript jQuery SQL Visual Basic Linux Bash FileMaker Script C

Key Takeaways

- **mBERT wins on performance**
  - The fine-tuned model beats the rule-based and semantic systems by a wide margin, achieving exact **F1 Scores** up to **0.6**.
  - It avoids the expensive manual work to keep taxonomies updated.
- **Occupations are easy, skills are hard**
  - Candidate's job titles are easy to find (with partial F1-scores up to 0.8).
  - Figuring out exactly **where** a **skill begins** and ends is still **very hard** for AI because people describe skills in different ways.
- **Resumes are much harder than job ads**
  - Job ads are usually clean and organized.
  - **Resumes** can be personal, and written in many **different styles**.
  - Our system found roughly 50% of the skills in English job offers, but that success rate dropped by half (to just 26.6%) on English resumes.

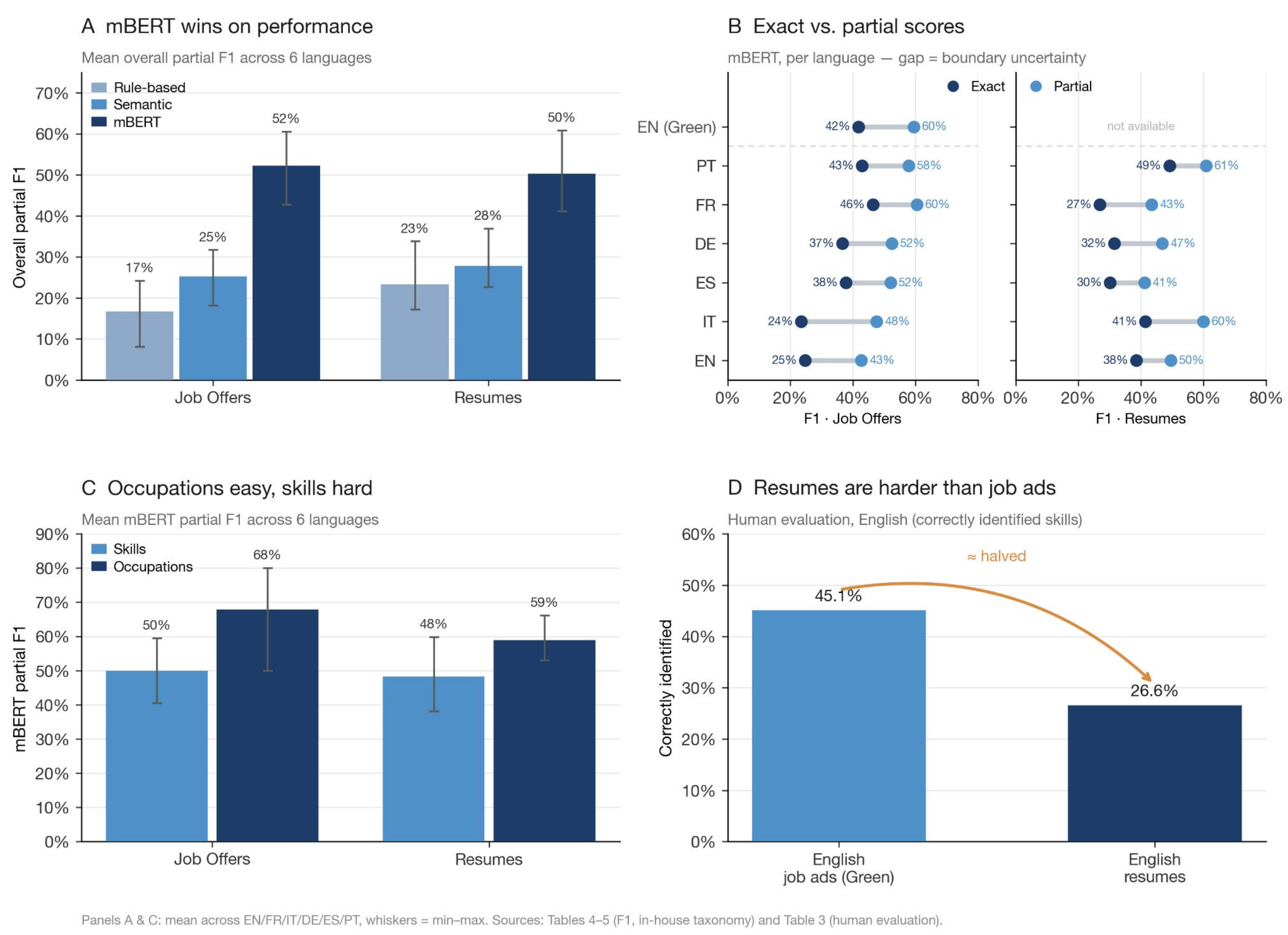

Figure 2. Skill Extraction Systems Results

*Work done while at Idiap Research Institute