Hardware-effective Approaches for Skill Extraction in Job Offers and Resumes

Laura Vásquez-Rodríguez, Bertrand Audrin, Samuel Michel, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa, Lonneke van der Plas

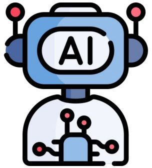
EHL
idiap RESEARCH INSTITUTE
ARCA 24 HR Tech Factory

RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources

1 Introduction

Skill Extraction Task Overview

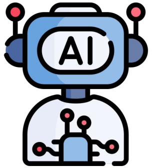
Skill Extraction Task



Data analysis using Python and Pandas

SKILL SKILL

Skill Extraction Task



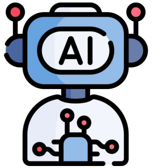
Data analysis using Python and Pandas

SKILL SKILL

Data analysis using Python and Pandas

SKILL SKILL SKILL

Skill Extraction Task



Data analysis using Python and Pandas

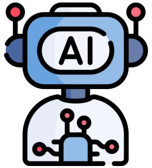
SKILL SKILL

Data analysis using Python and Pandas

SKILL SKILL SKILL

Data analysis using Python and Pandas

Skill Extraction Task



Data analysis using Python and Pandas

SKILL SKILL

Data analysis using Python and Pandas

SKILL SKILL SKILL

Data analysis using Python and Pandas

Scope: Skill Extraction

- Focus on hard skills and occupations
  - Soft skills for future work
- Task based on Named Entity Recognition (NER)
  - Span-based skill extraction
  - Straight-forward with existing datasets
- Evaluation exact and partial schema

2 Motivation

Hardware-effective Skill Extraction Task (Job offers and Resumes)

Our Motivation

- Understand current situation
  - Skill extraction in job offers and resumes
- Make the difference
  - Soft skills, bias, optimal hardware resources
  - Hardware requirements not a priority

Our Contributions

- A comparative evaluation of rule-based, semantic and neural models in skill extraction
  - Both job offers (public) and resumes (private)
- Hardware requirements analysis
  - Trade off performance and resources
- Qualitative evaluation of systems' outputs

3

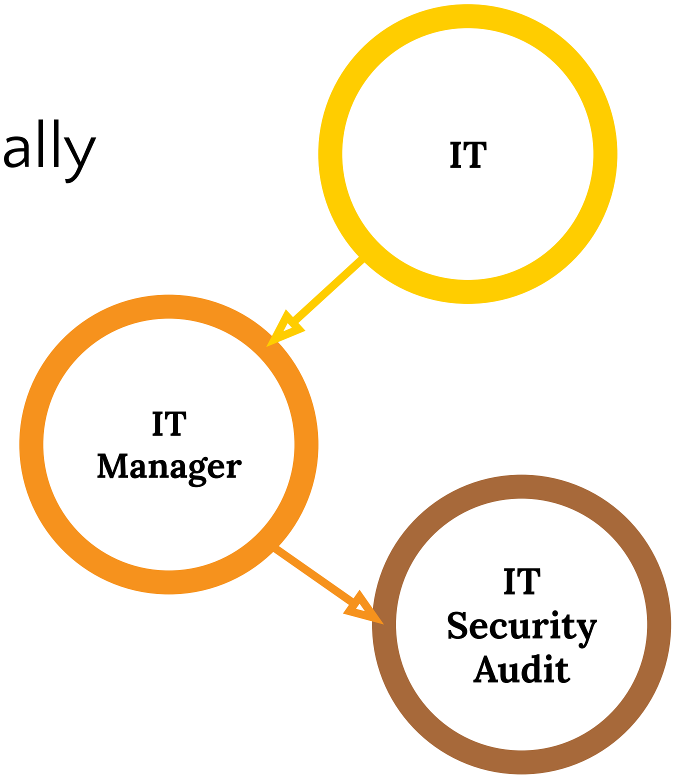
Comparative Evaluation

Rule-based, Semantic and Neural Models

Skills Taxonomy

- Collection of skills hierarchically connected to occupations

- Taxonomy:
  - ESCO
    - 131,623 entities, public
  - Arca24 company
    - 10,379 entities, private

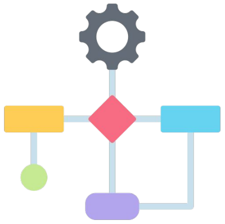


Methods in a Nutshell

Rule-based

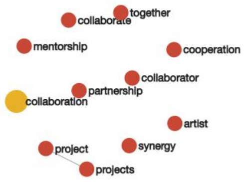
Keyword-based search, words standardization.

Python development -> Python develop



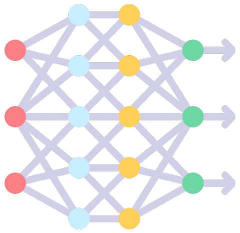
Semantic

Similarity (spacy) computation between taxonomy and texts spans in job offer, resumes.



Neural

Supervised BERT-based models trained on public and private annotated data (sentences)



Performance (exact, highest, F1)

	ArcaDB	ESCO	ArcaDB	ESCO
System	**Job Offers**		**Resumes**	
Rule-based	18.3%	20%	18,2%	14%
Semantic	12.9%	11.4%	8.6%	8.8%
BERT Supervised (no taxonomy)	45.9%		**51%**	

Green_JOB (8,669, 964, 335)
Arca24_CV (1,564, 195, 195)

Performance (exact, highest, F1)

	ArcaDB	ESCO	ArcaDB	ESCO
System	**Job Offers**		**Resumes**	
Rule-based	18.3%	20%	18,2%	14%
Semantic	12.9%	11.4%	8.6%	8.8%
BERT Supervised (no taxonomy)	45.9%		**51%**	

Performance (exact, highest, F1)

	ArcaDB	ESCO	ArcaDB	ESCO
System	**Job Offers**		**Resumes**	
Rule-based	18.3%	20%	18.2%	14%
Semantic	12.9%	11.4%	8.6%	8.8%
BERT Supervised (no taxonomy)	45.9%		**51%**	

Performance (partial, highest, F1)

	ArcaDB	ESCO	ArcaDB	ESCO
System	**Job Offers**		**Resumes**	
Rule-based	31.2%	39.8%	30.5%	26.9%
Semantic	24.8%	32.4%	24.5%	24.2%
BERT Supervised (no taxonomy)	66%		59.8%	

Performance (partial, highest, F1)

	ArcaDB	ESCO	ArcaDB	ESCO
System	**Job Offers**		**Resumes**	
Rule-based	31.2%	39.8%	30.5%	26.9%
Semantic	24.8%	32.4%	24.5%	24.2%
BERT Supervised (no taxonomy)	66%		59.8%	

4 Hardware Analysis

Trade-off between Resources and Performance

Hardware (HW) Analysis: Report

- Total time (m/h):
  - Preprocessing / training and test time per model
- CPU memory used (8GB)
  - AMD EPYC 7742 64-Core Processor
- GPU memory used (24GB)
  - NVIDIA GeForce RTX 3090

HW Analysis: Rule-based / Semantic

- Rule-based and semantic use less than 5GB RAM
  - They are solely run in CPU
  - Explainability, why a concept was chosen?
- Taxonomies comparisons can take long, but methods can still be improved

HW Analysis: Neural

- BERT and JobBERT models (~100M) use less than 8GB RAM (CPU)

- Although we do use GPUs..
  - Total time is less than 7 min
  - Using less than 9GB GPU RAM

HW Analysis: Neural

What about ESCOXLM-R?[1] (559M)

- Model is 5X bigger
- It uses about 3-4X more resources
- It showed less than 1% improvement (F1)

- More data? Limits of the skill annotation task?

[1]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.

5 Qualitative Analysis

Systems outputs & error analysis

Evaluation: System Outputs

Rule-based, ESCO_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Semantic, Arca_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Supervised, Green

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Evaluation: System Outputs

Rule-based, ESCO_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Semantic, Arca_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Supervised, Green

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Evaluation: System Outputs

Rule-based, ESCO_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Semantic, Arca_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Supervised, Green

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Evaluation: System Outputs

Rule-based, ESCO_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Semantic, Arca_DB

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Supervised, Green

The Test Consultant Automation Test Analyst will ideally be confident with Selenium and good experience of web-based testing HTML and Javascript

Evaluation: Error Analysis

- 120 random sentences in total
  - *bert-based-cased* (best)
  - *escoxlmr_skill_extraction* (larger)

- Classify sentences according to categories proposed by Nguyen et al., 2024[1]

[1]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).

Evaluation: Error Analysis

1. Skill definition misalignment
2. Wrong extraction
3. Conjoined skills
4. Extended span
5. Incorrect annotations
6. Others, e.g., partial detection

Evaluation: Error Analysis

- Category #1, skill definition misalignment (~50%)
- Category #2, wrong extraction (16.76%)
- The remaining sentences are distributed along the other categories

Do not underestimate the power of simple solutions. They can be essential to shorten the gap between research and the application

“

SEM24 project is funded by Innosuisse (grant 104.069 IP-ICT)

Thanks!

Any **questions** ?

You can find me at

- Laura Vásquez-Rodríguez
  - laura.vasquez@idiap.ch

Paper

LinkedIn







Icons in this presentation are credited to www.flaticon.com. Presentation template by SlidesCarnival