Extraction of Skills and Benefits from Job Postings Descriptions

NI-MVI Semestral Work Milestone

Matus Botek, botekmat@fit.cvut.cz

1.12.2023

1 Assignment

The goal of my project is to develop a method for extracting relevant skills and benefits from job posting descriptions. I should also focus on performance optimization. The validation should be done on a sample from the dataset that will be manually annotated by me. I will be using Kaggle's Indeed Job Posting Dataset[1].

2 Research

This section summarizes the articles and studies that I have read and highlights the most important outcomes.

2.1 A Survey on Skill Identification From Online Job Ads[1]

The first paper I have read gives an extensive overview of the state-of-the art approaches to online job ads skills extraction. I found multiple papers cited here that I followed later on. It lists available skill bases and divides them into categories, describes multiple approaches to skill identification and showcases recent applications of mentioned methods for the job application related tasks.

In the section that focuses on skill bases, ESCO seemed to be a good choice as it contained the most information compared to other options and it is multi-lingual. I have also seen it utilized in multiple papers that proposed methods for skill extraction.

[1] https://www.kaggle.com/datasets/promptcloud/indeed-job-posting-dataset

The paper showed the trends of usage for different skill extraction methods over the recent years. The simplest method was skill count, followed by topic modeling, skill embeddings and ML-based methods. Skill embeddings methods used word2vec or it's extensions (FastText).

For ML-based methods, the paper mentioned using Named Entity Recognition (NER) with LSTMs. Some works also focused first on a classification task, where they detected sentences that contain skills and in the next step, extracted those skills. One study compared CNNs and LSMT models for the sentence classification. Another approach was fine-tuning a pre-trained BERT model for the sentence classification. The last one was using a multi-label text classification for labeling each job description with the skills within, also leveraging BERT.

The conclusion also addressed possible paths that could be taken in this area such as using Graph Embeddings, creation of better, open job datasets and skill bases.

2.2 Language Model Based Extreme Multi-label Classification Framework[2]

In this paper they used the multi-label classification approach. They also created a dataset from Singaporean job ads and annotated the skills within each job posting. I downloaded the dataset and checked a couple of records. The job description was created by concatenation of "roles and responsibilities" and "job requirements" fields from the job postings and each records has a list of existing skills as labels. Their

architecture uses a pre-trained BERT model whose output is fed to a bottleneck layer that compresses the encoded representation that is then fed to a fully connected layer with sigmoid activation that does the multi-labelling. Their code is available on Github[2]

2.3 Data Pipeline for Labor Market Skills Extraction and Matching[3]

This paper proposed a DataOps pipeline for transforming big data into knowledge. They described their process of data collection, preprocessing, transformation, data mining / machine learning and interpretation in general. They showcased it on sentence classification using a simple neural network versus a pre-trained BERT model. It seems they only predicted if a sentence does or does not contain a skill, even though the paper talks about skill extraction with matching again an ontology.

2.4 An end-to-end framework for information extraction from Italian resumes[4]

This paper worked with CVs instead of job postings, but I have read it because they also extracted skills, but from CVs. First they extracted data from the resumes, divided them into thematic segments (personal information, education, skills, work experience) and performed a NER task on each segment. Their solution used a pre-trained BERT Italian model with a token-wise classifier on top. For evaluation they used recall, precision and F1-score. Later they compared the impact of different embedding approaches to the NER task, using BERT, FastText and 3 other models based on recent papers, with BERT staying on top.

2.5 A Context-Aware Approach for Extracting Hard and Soft Skills[5]

This paper showed an interesting approach to the task of skill extraction. They used EMSCAD dataset and transformed each sentence into the format of: left context, skill candidate (n-gram of up to 4 words), right context and a label (soft skill / hard skill / not skill). Then they utilized different word embedding representations with the combination of various ML classifiers. BERT embeddings enriched with part-of-speech and dependency parsing tags combined with logistic regression classifier performed the best for their sequence classification task.

2.6 Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads[6]

This paper used a Swiss-German job postings data. They first extract coarser spans containing skills (education, experiences, language). For this task, 2000 ads were iteratively annotated, treated as a NER task using spaCy and jobBERT-de (domain fine-tuned BERT model).

Then, these spans were used for fine-grained extraction of areas within those spans to give more detailed information for the data.

For mapping the extracted skills to ESCO concepts, semantic similarity lookup was used. To mitigate overly generic matches, they embedded contextualized spans and they also contextualized ontology terms as a term and it's class.

They also experimented with improving the vector similarity by different fine-tuning approaches to BERT. A masked language modeling on domain data approach seemed interesting to try. In the end a model that leveraged also other more advanced fine-tuning approaches showed to be the best.

2.7 Occupational skills extraction with FinBERT[7]

This is a Finnish master thesis focused on skill extraction using finBERT-cased model. I did not delve very deep into it, but the author manually annotated a subset of job ads using B-SKILL and C-SKILL tags, then used finBERT as a backbone model with a dense layer on top responsible for the NER task.

[2] https://github.com/WING-NUS/JD2Skills-BERT-XMLC

2.8 SkillSpan: Hard and Soft Skill Extraction from English Job Postings [8]

There are not many datasets openly released for the skill extraction task with suitable size, labels and description of the labeling process. This paper creates the dataset named SkillSpan. It was collected for over a year by scraping websites and part of it was annotated on span level. Based on ESCO rules, they distinguish between skills and knowledge. A detailed annotation guidelines were also released with it.

They also propose strong baselines using state-of-the-art language models with the focus also on the comparison of single-task vs. multi-task learning in this context. Mapping of extracted skills to a known ontology was not in the scope of this paper and could be subject to future enhancement.

As for the state-of-the-art models, they used and compared BERT, SpanBERT, JobBERT, JobSpanBERT with job variants being the respective models further pre-trained on domain-specific data (unlabeled SkillSpan data). The domain-specific pre-trained model is available on Huggingface[3]. All models used a CRF-layer and the experiments were carried out with MACHAMP, a toolkit for multi-task learning in NLP.

They also compared a sentence by sentence to whole job posting approach, utilizing Longformer model for the whole JP representation. The sentence by sentence approach showed better results. The continuos domain-specific pre-training also showed improvement compared to the non-adaptive models. Code for this paper together with the dataset is available on Github[4].

2.9 Skill Extraction from Job Postings using Weak Supervision[9]

I found this paper to be the quite useful. They extracted skills from job ads on span level and used ESCO taxonomy to find similar skills in the embeddings space. SkillSpan and Sayfullina datasets were used for this task, they considered both skills and knowledge labels just as skills to simplify them. The baseline used only some type of text vs skill matching without the use of a language model. Then, they created n-grams from the datasets and then compared them to the ESCO skill embeddings. Their solutions used 3 approaches for ESCO skills embeddings and used them with RoBERTa and JobBERT model. Here they also mentioned MACHAMP.

2.10 Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning[10]

I did not fully read this paper yet, but from what I understand so far, it is similar to other papers by Mike Zhang[8][9]. Here they release a Danish job dataset and make use of ESCO taxonomy as labels via distant supervision. I would read further into it but only after I try my initial approach that I summarized below.

3 My approach

As of now, I have spent most of my time doing the research and don't have a lot of code and no experiments done. So far my plan is to pre-process the data. While exploring my dataset I noticed a number of job ad duplicates, which I will remove. I want to clean all of the job descriptions from html tags using BeautifulSoup and new lines, extra whitespaces, etc.

For my first approach I want to try using the BERT architecture to create embeddings of the job descriptions on sentence level and compute similarity with embeddings of skills from ESCO as ontology. I have learned that a lot of the papers separate skill into multiple categories (soft skills / hard skills, explicit / implicit skills, skills / knowledge). My plan is to simplify this and just consider all of the mentioned categories as one skill category. I would also want to compare BERT with JobBERT model which was already pre-trained on the job domain and is avail-

[3] https://huggingface.co/jjzha/jobbert-base-cased
[4] https://github.com/kris927b/SkillSpan

able on Huggingface[5], ready for the skills extraction task. Maybe also pre-training an already pre-trained BERT model on my own using domain-specific data.

I also found a very recent paper published in July 2023 by the authors of [9] where they create a model called ESCOXLM-R (available on Huggingface[6] for skill extraction, pre-trained using ESCO and based on XLM-R model[11]. I did not read the article yet, but it also might be a good model choice for my task.

Overall, my planned initial approach is similar to the weak supervision approach mentioned in [9]. I want to simultaneously work on labeling a random subset of the data so that I can evaluate my methods, using embeddings and cosine similarity for the extraction and evaluate this approach on the manually-annotated data.

References

1. KHAOUJA, Imane; KASSOU, Ismail; GHOGHO, Mounir. A Survey on Skill Identification From Online Job Ads. *IEEE Access*. 2021, vol. 9, pp. 118134–118153. Available from DOI: 10.1109/ACCESS.2021.3106120.
2. BHOLA, Akshay; HALDER, Kishaloy; PRASAD, Animesh; KAN, Min-Yen. Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. In: SCOTT, Donia; BEL, Nuria; ZONG, Chengqing (eds.). *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 5832–5842. Available from DOI: 10.18653/v1/2020.coling-main.513.
3. GARRIGA, Martin; TAMBURRI, Damian; HEUVEL, Willem-Jan. DataOps for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction and Matching. In: 2020. Available from DOI: 10.1109/IRI49571.2020.00063.
4. BARDUCCI, Alessandro; IANNACCONE, Simone; LA GATTA, Valerio; MOSCATO, Vincenzo; SPERLÌ, Giancarlo; ZAVOTA, Sergio. An end-to-end framework for information extraction from Italian resumes. *Expert Systems with Applications*. 2022, vol. 210, p. 118487. ISSN 0957-4174. Available from DOI: https://doi.org/10.1016/j.eswa.2022.118487.
5. WINGS, Ivo; NANDA, Rohan; ADEBAYO, Kolawole John. A Context-Aware Approach for Extracting Hard and Soft Skills. *Procedia Computer Science*. 2021, vol. 193, pp. 163–172. ISSN 1877-0509. Available from DOI: https://doi.org/10.1016/j.procs.2021.10.016. 10th International Young Scientists Conference in Computational Science, YSC2021, 28 June – 2 July, 2021.
6. GNEHM, Ann-sophie; BÜHLMANN, Eva; BUCHS, Helen; CLEMATIDE, Simon. Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads. In: BAMMAN, David; HOVY, Dirk; JURGENS, David; KEITH, Katherine; O'CONNOR, Brendan; VOLKOVA, Svitlana (eds.). *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*. Abu Dhabi, UAE: Association for Computational Linguistics, 2022, pp. 14–24. Available from DOI: 10.18653/v1/2022.nlpcss-1.2.
7. CHERNOVA, Mariia. Occupational skills extraction with FinBERT. 2020.
8. ZHANG, Mike; JENSEN, Kristian Nørgaard; SONNIKS, Sif Dam; PLANK, Barbara. *SkillSpan: Hard and Soft Skill Extraction from English Job Postings*. 2022. Available from arXiv: 2204.12811 [cs.CL].
9. ZHANG, Mike; JENSEN, Kristian Nørgaard; GOOT, Rob van der; PLANK, Barbara. *Skill Extraction from Job Postings using Weak Supervision*. 2022. Available from arXiv: 2209.08071 [cs.CL].

[5] https://huggingface.co/jjzha/jobbert_skill_extraction

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

10. ZHANG, Mike; JENSEN, Kristian Nørgaard; PLANK, Barbara. Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning. In: CALZOLARI, Nicoletta; BÉCHET, Frédéric; BLACHE, Philippe; CHOUKRI, Khalid; CIERI, Christopher; DECLERCK, Thierry; GOGGI, Sara; ISAHARA, Hitoshi; MAEGAARD, Bente; MARIANI, Joseph; MAZO, Hélène; ODIJK, Jan; PIPERIDIS, Stelios (eds.). *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 436–447. Available also from: https://aclanthology.org/2022.lrec-1.46.

11. ZHANG, Mike; GOOT, Rob van der; PLANK, Barbara. ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain. In: ROGERS, Anna; BOYD-GRABER, Jordan; OKAZAKI, Naoaki (eds.). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 11871–11890. Available from DOI: 10.18653/v1/2023.acl-long.662.