Skills-Hunter: Adapting Large Language Models to the Labour Market for Skills Extraction

Antonio Serino[1,*,†]

[1]Department of Economics, Management and Statistics, University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126, Milan (MI), Italy

Abstract

In recent years, natural language processing (NLP) technologies have made a significant contribution in addressing a number of labour market tasks. One of the most interesting challenges is the automatic extraction of competences from unstructured texts. This paper presents an automated approach that exploits the latest NLP technologies, such as LLM, to overcome the data labelling problem in the task of extracting skills from job advertisements.

Keywords

Natural Language Processing, Large Language Models, Skill Extraction

1. Introduction

In the global economy, the labor market serves as the domain where the interplay between employer demand and applicant supply is characterized, forming a dense and intense system for the exchange of human resources. Over time, a series of social dynamics has caused this area to evolve more and more into the infosphere: the advent of job search platforms, such as Linkedin, has enabled the digitisation of the entire labour market, thus facilitating the emergence of Labour Market Information. Digital processing of data, including job adverts and CVs, enables the opportunity to collect, analyse and interpret a wide range of labour market data, including skills in demand. The analysis of large amounts of textual data allows the extraction of structured information from unstructured data. Skill extraction is still an open challenge that plays a key role in the labour market[1]. This task carries substantial importance as its successful completion would facilitate valuable analyses for various stakeholders in the sector. For instance, the effective extraction of skills from CVs could enhance the efficiency of selection processes. Furthermore, analyzing the trends of the skills most in demand in job advertisements could provide valuable insights for training institutions to better tailor their curricula.

In order to tackle all the tasks described, we need a tool that represents and organizes the skills and competences in an organized manner, providing a *lingua franca* for the labour market. ESCO

AIxIA'23: 22nd International Conference of the Italian Association for Artificial Intelligence, November 06–09, 2023, Rome, Italy

a.serino3@campus.unimib.it (A. Serino)

0009-0008-0737-8547 (A. Serino)

[1]https://www.etf.europa.eu/en/news-and-events/news/qualifications-and-skills-recognition-increasing-opportunities-building-fairer

is the multilingual European classification of Skills, Competences and Occupations, which aims to create a more knowledgeable and aligned European labour market by improving the match between labour demand and supply on the European labour market, facilitating labour mobility and human resource management on a transnational level, providing useful information to training providers, and keeping up-to-date through Data Driven approaches.

1.1. Research Objective and Question

Recent works in the literature are divided into that which exploits classification models that exploit annotated data to treat the task as a Named Entity Recognition (NER) problem and that which uses Large Language Models (LLM) to treat the extraction of skills as a text summary task. In the past year, LLMs have demonstrated remarkable flexibility, allowing the seampless execution of several tasks starting from the same pre-trained model without sacrificing good performance [1].

Our objective is to effectively enhance Large Language Models (LLMs) and tailor them to the domain of the labor market. This adaptation enables us to address critical tasks, such as skill extraction, while harnessing the pre-existing knowledge encapsulated within LLMs. This approach eliminates the necessity for resource-intensive and costly retraining of the whole model, while consistently achieving State-of-the-Art (SOTA) performance levels.

2. Background and Related Work

In this section, we will initially provide an exposition on two foundational technologies that underpin the proposed methodologies for skills extraction: Transformer models and Large Language Models (LLMs). Following this introduction, we will proceed to present relevant prior research related to automatic skills extraction.

2.1. Transformer Models

Transformer models [2] are a type of deep learning architecture that has made significant advancements in various other fields of machine learning, especially in Natural Language Processing (NLP). The key innovation of Transformer models is the use of a self-attention mechanism, which allows them to capture dependencies between different words or elements in a sequence, regardless of their distance from each other. This self-attention mechanism allows Transformers to excel at handling long-range dependencies and has largely replaced recurrent neural networks (RNNs) and convolutional neural networks (CNNs) in many NLP tasks. The self-attention mechanism allows the model to weigh the importance of different words or elements in a sequence when processing each word, capturing contextual information effectively. This, among others, allows to generate vector representations of words contextualized in the sentences they belong to.

In addition to the advantage of being able to realise multiple tasks through deep textual understanding, Transformer architectures can parallelise the computation of self-attention due to the inherent structure of the architecture that does not require sequential dependencies

between tokens. This has enabled training on huge amounts of data using devices with limited resources.

2.2. Large Language Models

Over the past year, the entire NLP has been revolutionised by the advent of Large Language Models, i.e. language models based on Transformer architectures whose number of parameters is exceptionally large.
LLMs are models based on transformer architectures: given an input sequence $X=(x_1,...,x_{i-1})$ where X is a text, an LLM is trained to predict the most probable token x_i, following x_{i-1}. To do so, they always exploit the self-attention mechanism, but extend the number of layers, as it has been shown that increasing the scale of parameters increases the models capabilities [3].

The increasing comprehension capabilities of the models make them capable of generating text. The generation of text by Large Language Models is a process based on conditional probability. The model uses an initial context, generally a sequence of tokens, to predict the probability of the next token in the sequence. This prediction is done through the Softmax formula, which converts the scores associated with each token into a valid probability distribution. The Softmax normalises the scores so that they represent a normality distribution. Then the model selects the next token based on the predicted probabilities, generating the one with the highest probability and continuing the generation process until it reaches a stop condition.

The ability to generate texts from a natural language input prompt, coupled with the capacity for deep understanding of texts, has enabled the emergence of conversational agents capable of performing certain tasks simply by using their pre-trained knowledge [4].

2.3. Related Works

As indicated above, various approaches have been explored in the recent literature to address the skills recognition task. One approach is the use of Transformer-based models to address the problem of competence recognition in a supervised learning contexts [5, 6]. However, these efforts have revealed a significant limitation, namely the scarcity of adequate annotated data. Often, too limited data samples are used, which do not capture the full range of existing variations in skills. Furthermore, the use of annotated data requires constant maintenance and updating in order to reflect emerging skills in the labour market. Another strategy investigated consists of adapting masked language models to the labour market context, using the ESCO taxonomy [7] as a reference, to perform various tasks, including skills recognition. A further attempt was to exploit the knowledge base of LLMs to tackle the skills recognition task through summarising and embedding strategiess [8]. However, this methodology could lead to a potential loss of information, especially since the summary is treated as a single embedding vector, thus compromising the complete representation of skills.

3. Proposed Method

As described in Sec.2.3, one of the main limitations of current skill extraction techniques, is the necessity of huge amounts of labelled data. To alleviate this problem, the first step of the

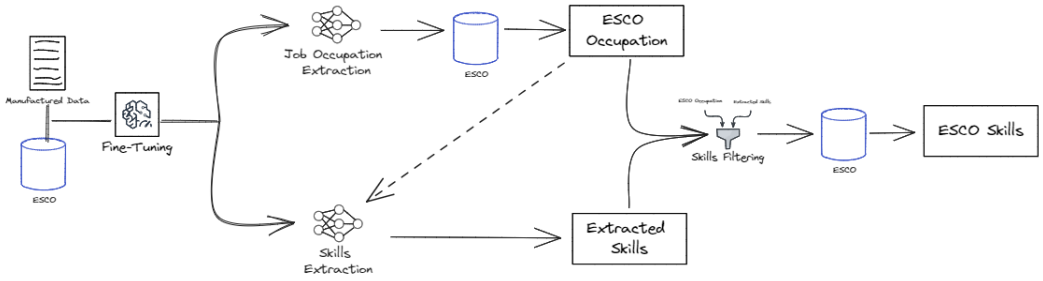

Figure 1: Proposed Pipeline

proposed methodology aims at generating realistic job vacancies with labels of the related occupation and the contained skills. Preliminary experiments have shown us that LLMs are able to generate Online Job Advertisements (OJAs) very close to the real ones, through the specification of the ESCO job profession and the ESCO skills that must be present in the OJA, being able to create a data structure based on key-value pairs that for each generated advertisement keeps track of the job title, OJA text and the skills present in the text. To improve the quality of such manufactured OJAs, we are going to fine tune them based on the output of a discriminator network, trained to discriminate real and manufactured data in a GAN fashion.

In the second step, we want to find a way that allows us to smartly merge manufacturing data with ESCO concepts, thus creating a data structure to fine-tune the model to fit its weights to the labour market domain and to fulfil two specific tasks: the recognition of job occupations described in the OJA and the extraction of explicit and implicit skills expressed in them. However, full fine-tuning presents some limitations: (i) first, it can lead to the so-called catastrophic forgetting [9], that is when a model trained on one task or dataset forgets its ability to perform well on a previously learned task or dataset when it is trained on a new and different task, and (ii) second, it require massive computational capabilities, especially when dealing with transformer models. Therefore, we resort to Parameter-Efficient Fine-Tuning (PEFT), a technique designed to fine-tune models while minimising the need for resources and high costs [10]. One of the possibilities of performing PEFT is Low-Rank Adaptation (LoRA). LoRA is an application methodology of PEFT that freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the architecture, greatly reducing the number of trainable parameters [11]. We will divide the manufacturing OJAs into two parts, using the first part to fine-tune the model and the second part to evaluate it.

We will create a first prompt that will take as a input an OJA asking the fine tuned model to process it and extract the recognised occupation label from the input OJA text. The recognised occupation label will be used as a query within a store containing the vector representation of all ESCO occupation labels and will return the most similar ESCO occupation, which we will define as OCC. We will then create a second prompt which will again contain the OJA provided as input and which will also specify the job occupation of the OJA. The objective of the second prompt is to ask the fine tuned model to parse the OJA and extract the explicit and implicit skill labels within it, relevant to the ESCO job occupation provided as input, adapting the skills to the semantic context described by the OJA to obtain a 'precise' extraction and asking it to

return the list of skills.

Having obtained the list of skills $S = (s_1, \dots, s_n)$ for each of its elements, we will perform a semantic comparison with the previously extracted ESCO occupation. We expect that, within a multidimensional vector space, the representation of relevant and pertinent skills for the input OJA is "close" to the representation of the job occupation described by the OJA. This would allow us to validate the semantic quality of the skills extracted by the model: all skills that exceed a similarity threshold will therefore be kept. This choice is also justified by the desire to try to overcome the possible problem of hallucinations [12], whereby the model might extract elements that are not semantically consistent with the OJA.

Having obtained the filtered list of skills $S = (S_1, \dots, S_m)$ with $m \leq n$, each of its elements will be used as a query within a second vector archive containing the representation of each ESCO skill, which will return as a result the top-1 skill closest to the skill used as a query. At the end of this, given an OJA, we expect to obtain as output the list of ESCO competences present within it. Having for each manufactured OJA the list of ESCO skills used for generation, we will perform an evaluation of the methodology by comparing the ESCO skills found and the ESCO skills effectively present, calculating the typical information retrieval metrics of Precision and Recall. Some preliminary experiments seem to overcome all the limitations present in recent SOTA work.

4. Expected Contributions to the Community

We proposed a new method for extracting and standardising skills based on the ESCO taxonomy. The idea is to adapt a model for the extraction of explicit and implicit skills. Current techniques in the area of Skills Intelligence suffers from the scarcity of labelled data. In this research we are going to study the automatic creation of a gold standard for the trainig and validation of skill extraction methods. Then, we proposed a skill extraction methodology using LLMs. In addition, another interesting application could be to create a prompter for new or alternative skills labels, useful for enriching the ESCO taxonomy. Future developments could be the development of a tool for filtering OJA, with the aim of reducing noise and optimising calculation time by keeping only those parts of the OJA relevant for competence extraction.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv (2020).

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023.

[5] A. Nigam, S. Tyagi, K. Tyagi, A. Saxena, Skillbert:"skilling" the bert to classify skills! (2020).
[6] M. Chernova, Occupational skills extraction with finbert, 2020.
[7] M. Zhang, R. van der Goot, B. Plank, Escoxlm-r: Multilingual taxonomy-driven pre-training for the job market domain, arXiv (2023).
[8] N. Li, B. Kang, T. De Bie, Skillgpt: a restful api service for skill extraction and standardization using a large language model, arXiv (2023).
[9] V. V. Ramasesh, A. Lewkowycz, E. Dyer, Effect of scale on catastrophic forgetting in neural networks, in: International Conference on Learning Representations, 2021.
[10] Z. Hu, Y. Lan, L. Wang, W. Xu, E.-P. Lim, R. K.-W. Lee, L. Bing, X. Xu, S. Poria, Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models, 2023.
[11] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2021.
[12] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. URL: https://doi.org/10.1145%2F3571730. doi:10.1145/3571730.