Extreme Multi-Label Skill Extraction Training using Large Language Models

Jens-Joris Decorte[1,2], Severine Verlinden[2], Jeroen Van Hautte[2],
Johannes Deleu[1], Chris Develder[1], and Thomas Demeester[1]

[1] Ghent University – imec, 9052 Gent, Belgium
{jensjoris.decorte, johannes.deleu, chris.develder,
thomas.demeester}@ugent.be
https://ugentt2k.github.io/
[2] TechWolf, 9000 Gent, Belgium
{jensjoris, severine, jeroen}@techwolf.ai
https://techwolf.ai/

Abstract. Online job ads serve as a valuable source of information for skill requirements, playing a crucial role in labor market analysis and e-recruitment processes. Since such ads are typically formatted in free text, natural language processing (NLP) technologies are required to automatically process them. We specifically focus on the task of detecting skills (mentioned literally, or implicitly described) and linking them to a large skill ontology, making it a challenging case of extreme multi-label classification (XMLC). Given that there is no sizable labeled (training) dataset are available for this specific XMLC task, we propose techniques to leverage general Large Language Models (LLMs). We describe a cost-effective approach to generate an accurate, fully synthetic labeled dataset for skill extraction, and present a contrastive learning strategy that proves effective in the task. Our results across three skill extraction benchmarks show a consistent increase of between 15 to 25 percentage points in *R-Precision@5* compared to previously published results that relied solely on distant supervision through literal matches.

Keywords: Skill Extraction · Contrastive Learning · LLM Generated Data.

1 Motivation and Related Work

Job ads are published online on a daily basis. They contain valuable information about economic trends in the labor market, such as the evolution of skill demand in time. Given that vacant jobs are advertized in unstructured text, we need automatic information extraction methods, e.g., to extract such mentioned skills. Such information extraction is crucial in labor market analysis and e-recruitment applications, including resume screening and job recommendation systems. Thus it is unsurprising that in the last decade, the number of studies on skill extraction methods has increased tenfold [6].

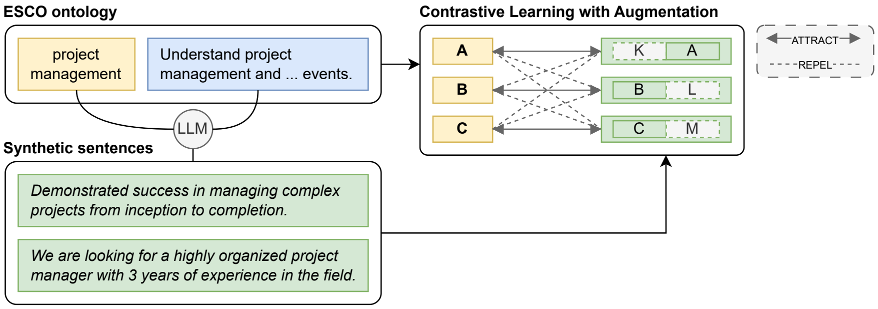
ESCO ontology
project management
Understand project management and ... events.
LLM
Synthetic sentences
Demonstrated success in managing complex projects from inception to completion.
We are looking for a highly organized project manager with 3 years of experience in the field.
Contrastive Learning with Augmentation
A
B
C
K
A
B
L
C
M
ATTRACT
REPEL

Fig. 1: High-level illustration of our methodology. We start from the ESCO ontology, which contains skills and their descriptions. We use a LLM to generate synthetic sentences. A bi-encoder is then optimized with contrastive learning to encode skill names and corresponding sentences in the same space. The sentences are augmented by randomly adding other sentences in front or behind them.

Several works have simplified the skill *extraction* problem to a pure *detection* task, limited to identifying the text span expressing a skill. Such a solution thus forgoes the normalization of synonyms and paraphrased skills toward an ontology of skill labels. Yet, since the latter is crucial for a consistent and robust job market analysis, skill extraction has been approached as an extreme multi-label classification (XMLC) task, requiring linking job ads to the relevant fine-grained skills out of a long list of skill labels in a given ontology [1,2,11,12]. These XMLC works suffer from the difficulty of constructing a qualitative training dataset: the huge number of labels (e.g., around 14k in our case study) makes annotation for skill extraction slow, and annotators are not sufficiently knowledgeable in each domain of the ontology to perform the task accurately.

In this paper, we present the effective usage of Large Language Models (LLMs) to circumvent the difficulty of annotation by automatically generating synthetic training data for skill extraction. The following paragraphs motivate our approach, which can be summarized as follows. We use an LLM to generate training data for skill extraction, grounded in the ESCO ontology. Based on this synthetic data, we optimize a model using contrastive learning to represent skill names and corresponding sentences close together in the same space. Our key contribution is a novel end-to-end approach to training a skill extraction system, consisting of the cost-effective synthetic data generation and the contrastive learning procedure alongside an effective augmentation procedure. The effectiveness of this method is compared against a distant supervision baseline on three skill extraction benchmarks. We release a large dataset of 138K *(skill, job ad sentence)* pairs, covering 99.5% of the ESCO ontology.

Large Language Models for skill extraction: With the increasing capabilities of LLMs over the past years, we argue that the time and knowledge-intensive task of annotation for skill extraction has become more feasible. The usage of LLMs for skill extraction has been proposed by SkillGPT [7], who use an LLM to summarize the skill requirements of job ads, after which the summary is embedded and compared with the ESCO skills through cosine similarity. While effective, as the authors themselves point out, this method suffers from the fact that generating the summaries is non-deterministic (implying that results would not be fully reproducible). Further, we note that needing to run each individual job ad at inference time through the LLM, also incurs non-negligible costs and latency. From the latter observation, we argue that a more effective way is to use the LLM only at training time, and in particular focus on constructing a strong training dataset, that far exceeds the quality of training data that has been constructed manually before. As illustrated in Fig. 1, we combine the domain knowledge captured in the ESCO ontology with the language understanding and generation capabilities of a large language model, to gather 138k pairs of skills and corresponding synthetic job ad sentences.[3]

Contrastive learning for skill extraction: The ESCO skill synonyms and descriptions have recently been used in a contrastive learning setup for skill extraction from German job ads [4]. The authors propose a two-stage approach for skill detection: first identifying skill mentions as text spans, followed by ranking the ESCO skills against those mentions. The ranking is performed by a bi-encoder model, which is optimized through contrastive training on pairs of *(skill, description)* and *(skill, synonyms)* from ESCO [4]. However, we believe this approach is sub-optimal for two reasons. First, skills are often mentioned as longer implicit spans or even full sentences, which pose challenges to span detection approaches. Second, ranking ESCO skills solely based on the detected spans restricts the utilization of contextual information and the strong contextual representation capabilities of BERT-based models. Based on these insights, we adopt a contrastive learning approach that operates directly on full sentences, by training a bi-encoder to represent sentences and their corresponding skill labels closely together in representation space. This strategy draws on a recent contribution demonstrating the strength of applying contrastive learning on pairs of biomedical concept names, and their corresponding textual descriptions [10].

2 Methodology

We make use of the ESCO ontology, which is briefly described in Section 2.1. Section 2.2 describes how the ESCO skills and their descriptions are presented to an LLM to generate a large-scale synthetic dataset of positive *(skill, job ad sentence)* pairs. Finally, this data is used in a contrastive learning procedure to optimize a bi-encoder for skill extraction, as detailed in Section 2.3.

2.1 Skill ontology: ESCO

For this work, we use version 1.1.0 of the *European Skills, Competences, Qualifications and Occupations* (ESCO) ontology [3]. ESCO contains names and descriptions of over 3k occupations and almost 14k skills, in 28 different languages.

[3] The data will be released upon acceptance.

We focus on the English version of ESCO and we only rely on skills. These skills have on average 7 additional English synonyms linked to them, but on first inspection, these are not always entirely accurate, and we decided not to use them for our experiments (although the proposed method can be applied directly to the skill names augmented with the synonyms). Each skill is furthermore accompanied by a textual description, on which our approach toward synthetic data generation strongly relies.

2.2 Synthetic Data Generation

Aiming for a contrastive learning strategy, we need a set of positive training pairs *(skill, job ad sentence)*, for all skills in the considered label space (i.e., ESCO skills). Rather than starting from sentences and prompting the LLM to generate skill labels (for which an iterative strategy such as [8] may be used), we start from a single ESCO skill and prompt the LLM to generate sentences *as if* from job ads, that require that skill. As such, we avoid the difficulty of aligning the LLM output with the ESCO skill label set. The *gpt-3.5-turbo-0301* model was used through the paid API provided by OpenAI. The model was chosen because of its intuitive prompting procedure and competitive cost.

LLM prompt design: Two prompts were compared (provided in Appendix A). The first prompt requests a list of job ad sentences requiring a certain skill, with demonstrations for the skills *Java* and *project management*. With this method, the LLM responded with a list of sentences for 73% of the ESCO skills. In the other cases, the response stated that not enough information about the skill was provided to generate accurate sentences, or that it would be highly unlikely that job ads request the skill. The second prompt was obtained by extending the initial prompt to include the ESCO skill description, as well as a clarification that the requested job ads are *hypothetical*. This increased the number of responses with lists of sentences from 73% to 99.5%. The accuracy of the generated data was manually assessed on a subset, and determined to be at around 88% and 94%, respectively, for the first and second prompts. The final dataset was generated with the second prompt and consists of 10 synthetic sentences for 13,826 unique ESCO skills. Some examples from this dataset are listed in Appendix B.

2.3 Contrastive Learning for Skill Extraction

For the contrastive learning, we use a bi-encoder architecture, in which pairs of skills and corresponding sentences are encoded by the same encoder-only transformer architecture [9]. We make use of *multiple negatives ranking loss* with in-batch negatives, as proposed by [5]. The setup is visualized in Fig. 1. Inspired by the work in [4], we also train a model using *(skill, ESCO description*) pairs, to compare the benefit of using the synthetic job ad sentences. The trained bi-encoder is directly used for the task of skill extraction, by ranking all ESCO skills with respect to an input sentence, based on the cosine similarity of their representations. In other words, no supervised fine-tuning is performed on the skill extraction task.

Augmentation: We note that the synthetic job ad sentences typically only discuss its linked skill, while real sentences often mention multiple skills or concepts. We hypothesize that this setup limits the model's capability to reflect multiple skills in its embeddings, which in turn would harm the performance for skill extraction. To this end, we introduce an augmentation strategy that randomly adds another sentence in front or behind each sentence during training. The pairs now consist of a skill and two concatenated sentences, of which only one sentence relates to the skill. We argue that this forces the model to represent both topics of the concatenated sentences to match with the correct skill in the batch. This augmentation is visualized in Fig. 1.

3 Experimental Setup and Results

We start from a sentence-transformer model (*all-mpnet-base-v2*[4]) that was pretrained on over 1B English sentence pairs [9]. The synthetic training dataset is compared to directly using pairs of ESCO skills and their descriptions, with and without the proposed augmentation method. For each dataset, we use the same hyperparameters, reported in Appendix C. Skill extraction performance is evaluated against the TECH and HOUSE benchmarks provided in [2], and an additional proprietary test set TECHWOLF, see Appendix D for details. Results are expressed in terms of mean reciprocal rank (MRR) and r-precision at 5 (RP@5), as in [2], shown in Table 1. RP@K is defined in (1), where the quantity $Rel(n, k)$ is a binary indicator of whether the k^{th} ranked label is a correct label for data sample n, and R_n is the number of gold labels for sample n.

$$RP@K = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\min(K, R_n)} \sum_{k=1}^{K} Rel(n, k) \tag{1}$$

	TECH		*HOUSE*		*TECHWOLF*	
Metric	MRR	RP@5	MRR	RP@5	MRR	RP@5
Distant Supervision [2]	32.50	32.12	29.59	30.57	28.44	29.27
all-mpnet-base-v2	38.76	39.60	26.27	26.17	29.58	33.48
ESCO descriptions	46.83	48.10	36.46	37.17	42.09	44.96
ESCO descriptions$^{\text{AUG}}$	48.46	51.99	39.15	42.35	44.13	45.87
GPT sentences	48.33	48.80	41.13	40.80	46.50	51.24
GPT sentences$^{\text{AUG}}$	**52.85**	**54.62**	**42.75**	**45.74**	**52.55**	**54.57**

$^{\text{AUG}}$: with proposed augmentation.

Table 1: Evaluation of different training data regimes. The distant supervision method from [2] is reproduced. The pretrained *all-mpnet-base-v2* model is also directly evaluated, without any contrastive learning. Best results are **bold**, second best results are underlined.

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Synthetic sentences yield superior performance to ESCO descriptions, and augmentation has a positive impact in all cases. We analyze the performance as we randomly subsample the number of synthetic sentences per skill. As shown in figure 2, more data helps, but even one synthetic sentence outperforms the description-based models.

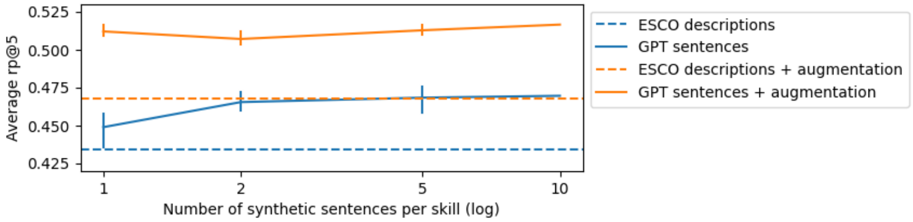

Fig. 2: Average rp@5 across all benchmarks, for different amounts of sentences per skill during training. Note that just using one sentence per skill already outperforms the description-based training. The horizontal axis is in logarithmic scale. Error bars show the minimum and maximum performance of the runs.

Note that the sentence representation of the model is simply an average of the embeddings of each token in the input. This allows us to understand the relation between the predicted skills and the words in the input. Appendix E contains a detailed explanation of how to do this, with some visual examples that help interpret the relation between skills and sentences in the model.

4 Conclusion

This paper presents a cost-effective method for generating a comprehensive synthetic dataset of sentences, grounded in the ESCO ontology. The size of this dataset surpasses any previously annotated dataset for skill extraction and covers 99.5% of skills in ESCO. We demonstrate that a bi-encoder can be optimized using a contrastive training procedure to effectively represent both skill names and corresponding sentences in close proximity within the same space. This approach outperforms our distance supervision baseline by a large margin. Additionally, we propose a simple augmentation method that enhances the resulting model quality. We release the full dataset to foster future research in this area.

Ethical considerations

The skill tagging task is frequently used within human resources. Algorithms can amplify or introduce new biases in this domain. Therefore, it is crucial to carefully monitor and reduce any potential biases that might emerge from the data generation procedure. Additionally, a constant classifier performance should be assured across different population groups. Finally, any development of a skill extraction method should keep in mind the final application for which the extracted skills will serve.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This project was funded by the Flemish Government, through Flanders Innovation & Entrepreneurship (VLAIO, project HBC.2020.2893).

References

1. Bhola, A., Halder, K., Prasad, A., Kan, M.Y.: Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5832–5842. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.513, https://aclanthology.org/2020.coling-main.513
2. Decorte, J.J., Van Hautte, J., Deleu, J., Develder, C., Demeester, T.: Design of negative sampling strategies for distantly supervised skill extraction. In: Proc. 2nd Workshop Recomm. Sys. Hum. Resour. at RecSys 2022 (RecSys in HR 2022). pp. 1–7. Seattle, WA, USA (22 Sep 2022)
3. ESCO: European skills, competences, qualifications and occupations. EC Directorate E (2017)
4. Gnehm, A.s., Bühlmann, E., Buchs, H., Clematide, S.: 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). pp. 14–24. Association for Computational Linguistics, Abu Dhabi, UAE (Nov 2022), https://aclanthology.org/2022.nlpcss-1.2
5. Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.H., Lukács, L., Guo, R., Kumar, S., Miklos, B., Kurzweil, R.: Efficient natural language response suggestion for smart reply. ArXiv **abs/1705.00652** (2017)
6. Khaouja, I., Kassou, I., Ghogho, M.: A survey on skill identification from online job ads. IEEE Access **9**, 118134–118153 (2021)
7. Li, N., Kang, B., Bie, T.D.: Skillgpt: a restful api service for skill extraction and standardization using a large language model (2023)
8. Raedt, M.D., Godin, F., Demeester, T., Develder, C.: IDAS: Intent discovery with abstractive summarization. In: The 5th Workshop on NLP for Conversational AI (NLP4ConvAI@ACL) (2023), https://arxiv.org/abs/2305.19783
9. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410
10. Remy, F., Demuynck, K., Demeester, T.: BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 1454–1465. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.findings-emnlp.104
11. Vermeer, N., Provatorova, V., Graus, D., Rajapakse, T., Mesbah, S.: Using RobBERT and eXtreme multi-label classification to extract implicit and explicit skills from Dutch job descriptions (2022)

12. Zhang, M., van der Goot, R., Plank, B.: ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain (2023)

Appendix A Prompts

We make use of the conversational "user" and "assistant" roles in the *gpt-3.5-turbo-0301* model to integrate demonstrations into the prompt. The information about the ESCO skill at hand needs to be pasted into the prompt, indicated by *variables* in the prompt. The first version of our prompt is shown below:

```
User: Write two sentences from job ads that require the skill Java.

Assistant: - experience with Java development, preferably web-based
- looking for a Java programmer this summer

User: Write two sentences from job ads that require the skill project
management.

Assistant: - successful project managers are able to manage multiple
tasks and deadlines simultaneously
- being able to effectively manage projects can give you valuable
experience and skills

User: Write 10 sentences from job ads that require the skill skill.
```

The second version of the prompt is more lengthy. It makes use of the "system" message to specify that the requested job ads are hypothetical. The demonstrations follow a more structured format, and contain the ESCO description. The response format of the "assistant" remains the same.

```
System: Respond with sentences from hypothetical job ads that require a
certain skill, as asked by the user.

User: Number of sentences: 2
Skill: Java
Definition: The techniques and principles {...} in Java.

Assistant: - experience with Java development, preferably web-based
- looking for a Java programmer this summer

User: Number of sentences: 2
Skill: project management
Definition: Understanding project management and {...} events.

Assistant: - successful project managers are able to manage multiple
tasks and deadlines simultaneously
- being able to effectively manage projects can give you valuable
experience and skills

User: Number of sentences: 10
Skill: skill
Definition: skill description
```

Appendix B Examples of Synthetic Data

We include the synthetic job ad sentences for two skills from ESCO. The following sentences were generated for the skill `interpret medical results` with description *"Interpret, integrate and apply results of diagnostic imaging, laboratory tests and other investigations as part of the assessment of the client, in consultation with other healthcare practitioners."*

1. Applicants must be able to accurately interpret laboratory tests to assist in clinical decision making.
2. We are currently seeking a professional who can accurately interpret medical results.
3. We are looking for an experienced healthcare professional who can effectively communicate with others to provide quality care based on medical results.
4. we are looking for a medical professional who can apply their expertise in interpreting complex medical results to improve patient outcomes.
5. The position requires an understanding of how to integrate and apply diagnostic results in clinical practice.
6. working experience in a clinical setting and interpreting medical test results is an advantage.
7. we are seeking a candidate who can accurately interpret medical results and provide appropriate treatment recommendations.
8. Must have experience interpreting medical results in a clinical setting.
9. the candidate we are looking for must have the ability to integrate results of diagnostic tests and other investigations.
10. excellent analytical and problem-solving skills are needed to be able to interpret medical results.

Similarly, for the skill `shunt inbound loads` with description *"Shunt inbound freight loads to and from railcars for inbound and outbound trains."*, the synthetic sentences are:

1. we are seeking a skilled laborer to shunt inbound loads efficiently and safely
2. Qualified candidates should possess a strong understanding of shunting techniques as well as experience operating shunting equipment.
3. we need someone who can shunt inbound loads quickly and accurately
4. must have experience in shunting inbound freight loads
5. Successful applicants will have a track record of accuracy and attention to detail while shunting inbound loads.
6. We need someone who can effectively and efficiently shunt inbound loads to minimize downtime.
7. the ability to shunt inbound loads is a must-have skill for this position
8. shunting inbound loads is a physically demanding job
9. We are looking for experienced operators with the ability to shunt inbound loads.
10. Shunt drivers must be able to lift heavy loads and follow safety protocols.

Appendix C Training details

The contrastive training is implemented using the popular SBERT implementation [9]. We keep the default value of 20 for the "scale" hyperparameter *alpha*. We always train for 1 epoch. The positive pairs are randomly shuffled into batches of 64. We use the AdamW optimizer with a learning rate of 2e-5 and a "WarmupLinear" learning rate schedule with a warmup period of 5% of the training data. Automatic mixed precision (AMP) was used to speed up training. All experiments where performed using an Nvidia T4 GPU.

Appendix D Benchmark details

This paper compares three different benchmarks. In table 2 you find metrics for the size of the datasets and the amount of skills tagged for each sentence. The HOUSE and TECH dataset are publicly accessible, whereas the TECHWOLF dataset refers to proprietary data held by the company. Both the HOUSE and the TECWHOLF dataset comprises sentences from various job postings, whereas the TECH dataset exclusively contains sentences from the StackOverflow job posting platform. Each dataset was manually annotated with the skills of the ESCO ontology.

Metrics	HOUSE	TECH	TECHWOLF
# sentences	323	405	326
Average # of skills per sentence	3.03	2.04	1.80
# unique skills	389	286	314

Table 2: The table shows the amount of labeled sentences, the average amount of skills tagged per sentence and the amount of unique skills for each dataset.

Appendix E Visualization of Predictions

Table 3 contains the top 3 predictions for two sentences that both contain multiple skill concepts. The sentence representation of the model is simply an average of the embeddings of each token in the input. This allows us to understand the relation between the predicted skills and the words in the input. Concretely, we visualize the cosine similarity between the skill embedding and the embeddings of each of the tokens in the sentence. This provides a visual interpretation of the relative importance of the input words with respect to the skills, which we find to be an interesting approach to understanding what the model has learned.

Label	**Visualization**
Coach IT teams with strong C++ skills	
C++	coach it teams with strong c + + skills
Microsoft Visual C++	coach it teams with strong c + + skills
coach employees	coach it teams with strong c + + skills
Onboard new colleagues and manage salaries	
introduce new employees	onboard new colleagues and manage salaries
determine salaries	onboard new colleagues and manage salaries
hire new personnel	onboard new colleagues and manage salaries

Table 3: Top skills for each sentence. For each skill, the sentence is visualized with the cosine similarity between the skill embedding and each token embedding. Higher cosine similarity is indicated by darker background color.