Taxonomy Enrichment: A Framework for Automatic Updates and Labor Market Analysis*

Alessia De Santo[1,*,†]

[1]Department of Economics, Management and Statistics DEMS, Univ. of Milano-Bicocca, Milan, Italy

Abstract

Hierarchical taxonomies serve as fundamental structures for reasoning with hierarchical concepts across various domains such as healthcare, finance, and economy. However, maintaining their relevance and accuracy is a labor-intensive and error-prone task, demanding experts to identify and revise novel concepts constantly. In this context, distributional semantics techniques offer a promising avenue by suggesting terms likely to be associated with existing concepts. In our study, we propose a method to enhance taxonomies by adding related terms using contextual word embedding as encoders. We introduce VESPATE (**VE**ctor **SPA**ce model for **T**axonomy **E**nrichment), a system designed to automatically expand any given hierarchical taxonomy with new terms using three generative models. Additionally, we integrate VESPATE with human validation to identify and select the most suitable terms for inclusion in the taxonomy. VESPATE was deployed within an EU project to enrich the official European Skill taxonomy, ESCO, with 40K+ digital terms gathered from the Web, aligning ESCO skills with current labor market needs. A total of 924 terms were selected through VESPATE, with 757 new terms subsequently validated by domain experts for inclusion in the digital skills taxonomy. Our framework, employing a pool of LLMs as encoders, helped us mitigate the limitations of the generative model, reducing the potential for errors and ensuring precise results in taxonomy enrichment. Additionally, the initial implementation of VESPATE consistently decreased the human effort required for the project. We evaluated the robustness of our system against a closed-world evaluation constructed using ESCO's hierarchy, achieving a 81% Positive Predictive Value (PPV) when combining all three models.

Keywords

Automated Taxonomy Enrichment, Labour Market Intelligence, Large Language Models, NLP

1. Introduction and Motivation

Hierarchical taxonomies are essential tools for organizing and reasoning with complex concepts across diverse domains such as healthcare, finance, and economics. They provide a structured way to categorize and relate concepts, enabling better decision-making and analysis. Because the target domain of specific taxonomy changes over time, taxonomies must be kept up to date so that newly introduced categories and hierarchical relationships can be properly integrated [1]. However, manually constructing and maintaining taxonomies is expensive and time-consuming due to its labor-intensive and domain-specific nature [2]. Here arises the importance of developing automated taxonomy enrichment methods. This work presents a production system that combines generative models with domain experts' knowledge for realizing an automated taxonomy enrichment process and its validation. To develop our method, we resort to contextual embedding. The latter, pre-trained on large-scale unlabeled corpora, achieve state-of-the-art performance on a wide range of natural language processing tasks, such as text classification, question answering, and text summarization, [3, 4]. For this reason, we decided to use LLMs as encoders.

This research is framed in the context of an ongoing EU-funded project entitled "Towards the European Web Intelligence Hub - European system for collection and analysis of online job advertisement data (WIH-OJA)", which aims to put Online Job Ads into official statistics on the labor market, by constructing

Doctoral Consortium at the 23rd International Conference of the Italian Association for Artificial Intelligence Bolzano, Italy, November 25-28, 2024

*This research activity is partially supported within the EU project entitled "Towards the European Web Intelligence Hub - European system for collection and analysis of online job advertisement data (WIH-OJA)"

*Corresponding author.
alessia.desanto@unimib.it (A. D. Santo)

an AI-based system that collects and analyses million Online Job Ads over Europe[1]. Within this project, this paper aims to create an automated pipeline to enrich the ESCO[2] digital skill taxonomy with up-to-date information collected from the Web.

ESCO and similar taxonomies play a crucial role in identifying necessary skills for specific jobs, facilitating accurate tracking of labor market changes. Policymakers rely on such data to plan future interventions. Establishing a system to update taxonomies with real-world data is essential, particularly in digital skills, given the constant growth of human knowledge and the daily emergence of new concepts. In this scenario, the contribution of this work is threefold:

1. VESPATE - **VEctor SPAce model for Taxonomy Enrichment**. We formalize and propose a taxonomy enrichment method, namely VESPATE, that exploits contextual word embedding as encoders to enrich hierarchical taxonomy with additional related terms. The method is taxonomy independent as it can be applied to any hierarchical structure of concepts;

2. **Real-world deployed application**. We apply VESPATE to the field of labor market to enrich the official ESCO taxonomy, by evaluating up to 40K+ digital-related terms. VESPATE has been deployed within the aforementioned EU research project, planning to enrich ESCO periodically.

3. **Closed-world evaluation and Human Contribution**. We construct a closed evaluation leveraging ESCO structure to evaluate VESPATE and validate the results obtained from the application with human experts belonging to the European Network of Regional Labor Market Monitoring[3].

2. Related work

Taxonomy Enrichment Recently, the automated enrichment of generic taxonomies has gained relevance. Vedula et al. [5] use word embeddings to find semantically similar concepts in the taxonomy. Then they use semantic and graph features, some from external sources, to find the potential parent-child relationship between existing concepts and new ones from Wikipedia categories. Manzoor et al. [6] model the implicit edge semantics to score the hyponymy relevance between node pairs. Aly et al. [7] adopt hyperbolic embedding to capture hierarchical lexical-semantic relations and find orphans (disconnected nodes) and outliers (child assigned to wrong parents) in a taxonomy. Shen et al. [8] present position-enhanced graph neural networks to encode the relative position of nodes. They use a set of <query, anchor> concepts generated from an existing hierarchy to train a model to predict the parent-child relationship between the anchor and the query. Other works have tried to address this problem using self-supervised constructing training sets from existing taxonomy [9, 10, 11]. Cheng et al. [12] propose an adaptively self-supervised user behavior-oriented product taxonomy expansion framework. The self-supervised generation strategy avoids inheriting the adverse problems in the existing taxonomy.

Labor Market Intelligence (LMI) Taxonomies play a pivotal role in the LMI realm, and several contributions underscore their relevance for monitoring, analyzing, and comprehending labor market shifts. To cite a few: Alabdulkareem et al. [13] delved into the relevance of skills within the US standard occupations taxonomy, O*NET. Giabelli et al. [14] introduced WETA. This domain-independent method utilizes distributional semantics and classification for automatic taxonomy alignment. They use it for bridging the Italian occupation taxonomy and ESCO. Malandri et al. [15] propose to use word embedding to refine taxonomies and test it on ESCO and European online ads. Arslan and Cruz [16] use BERTopic to automatically augment a given business taxonomy with additional concepts extracted from a corpus of online news documents.

[1]Since 2019, the project "Towards the European Web Intelligence Hub - European system for collection and analysis of online job advertisement data (WIH-OJA)" collected 300 million unique online job ads from 32 EU Countries.
[2]Is the official European multilingual classification of Skills, Competences, and Occupations.
[3]http://regionallabourmarketmonitoring.net/

3. Development of VESPATE

Our work centers on the enrichment of an existing taxonomy, with a fundamental assumption that the core taxonomy remains unchanged (see previous works such as [17],[1]). The enrichment process focuses on integrating new, fine-grained terms into the existing framework without altering the established hierarchical relationships among the original concepts. This approach ensures that the enriched taxonomy gains additional specificity and detail while preserving its foundational structure. To achieve this, VESPATE operates in three key steps:

1. **Encoding Step:** The process starts by using an embedding model to generate vector representations for each existing concept in the taxonomy that could serve as a parent, as well as for each new term. These vectors capture the semantic meaning of the concepts and terms, allowing the model to assess their relevance effectively

2. **Scoring Step:** A scoring function is then applied to compute a scoring matrix. Each entry in this matrix represents the similarity score between a concept from the core taxonomy and a new term, indicating how closely related they are. Common measures for vector similarity, such as cosine similarity - the one we chose, are typically employed at this stage.

3. **Matching Step:** For each new term, the algorithm identifies the most suitable parent concept in the existing taxonomy—the one that maximizes the similarity score with the new term. The enriched taxonomy is subsequently constructed by incorporating these new terms into the existing set of concepts, extending the hierarchical relationships with newly identified parent-child pairs.

While the described algorithm applies to a single encoding model, our approach involves combining multiple Large Language Models (LLMs). Specifically, we employed three distinct LLMs. Research in collaborative machine learning has demonstrated that using multiple models concurrently can enhance overall performance, in improving confidence in correct output [18, 19]. This strategy is based on the idea that models trained differently exhibit varying strengths and weaknesses, enabling them to complement and correct one another. As a result, the confidence and accuracy of the outcomes may be enhanced.

The next section presents the results of applying this approach to an existing taxonomy, specifically the ESCO Digital Collection. These results informed our decision to combine the models' outputs by focusing on consensus, meaning that we chose to retain only those matches where all three models agree.

4. VESPATE Validation: A real-world application to the labour market

4.1. Enriching ESCO Digital Taxonomy

This study is situated within the framework of an ongoing EU-funded project which seeks to integrate Online Job Ads into official labor market statistics by developing an AI-based system capable of collecting and analyzing millions of Online Job Ads across Europe. Within this project scope, our work aimed to establish an automated pipeline for enhancing the ESCO digital skill taxonomy with current information gathered from the web.

Data The project focused on the ESCO Digital Skills taxonomy, a part of ESCO Skill taxonomy that addresses only digital skills. In particular, we target its lower-level nodes, which include 89 specific digital skills. To enrich this taxonomy, we utilized a dataset comprising 40,561 digital terms, extracted by third parties from two main web platforms: GitHub and Stack Overflow. These platforms are critical repositories for programming resources and discussions within the computer science and ICT communities. After filtering the dataset to focus on terms with higher-than-average mentions both on Stack Overflow and Github, we narrowed it down to 4,215 terms deemed suitable for integration into the ESCO taxonomy. Therefore our input data were the digital terms extracted from the web, which came also with a brief description, and the ESCO digital taxonomy which contained the digital skills, their definitions and some use examples. All these data were used as input for our embedding models.

Models Used To implement VESPATE we used three pre-trained Large Language Models (LLMs). These models were selected based on their performance in the Massive Text Embedding Benchmark (MTEB) [20], which assesses model capabilities across various embedding tasks, including Semantic Textual Similarity (STS). The selected models, each with 335 million parameters, were:*mixedbread-ai/mxbai-embed-large-v1*, *w601sxs/b1ade-embed*, and *Labib11/MUG-B-1.67*. Selected as the top three *open source* pre-trained models available on June 2024. These LLMs served as encoders to process the digital terms and match them with the appropriate nodes in the ESCO taxonomy. As described in the previous sections, the matching process involved calculating cosine similarity between the vector representations of the new terms and the ESCO skills. The framework selected the best matches where all three models identified the same ESCO skill as the most suitable parent.

Validation and Results To confirm the efficacy of our method in this specific task, we created a baseline to evaluate VESPATE performance, constructed in a closed setting: ESCO upper levels. The ESCO digital collection is hierarchically organized, with each subsequent level providing further specification of the one above it. Thus, to create this evaluation scenario, we proposed mimicking the functionality of our framework on the higher level of the whole taxonomy; specifically, we aimed to enrich the second level of ESCO using its third level, one already present in the taxonomy. Then, we utilized the outcomes of this process for performance evaluation. First, we perform an individual evaluation of the three models used. For each LLM, we calculate the Positive Predictive Value (PPV). We obtained a PPV rate of 78% for both MUG-B-1.6 and b1ade-embed while mxbai-embed-large-v1 achieved 79%. Then we tested the intuition on combining model results, considering only the ones where all three models agreed. To check if the idea that models might correct each other was valid, we considered the cases in which they all produced an incorrect match on the same term (i.e., none of the models got the right match), this happened only 14.3% of the time, suggesting that the models make different mistakes. Then, we analyzed the performance in cases where all three models agreed on a match. Out of 246 ESCO skills, the three models had a common match in 220 cases. Of these 220 matches, 178 were correct, resulting in an 81% accuracy rate for matches agreed upon by all three models. Building on our closed-world results, we applied VESPATE to enrich the final level of the ESCO Digital Taxonomy, a subset of ESCO specifically focused on digital skills. The framework's matching process was first employed to align new terms with the existing taxonomy. To further streamline the process and reduce the need for extensive human intervention, we introduced an additional refinement step: only matches within the top similarity quartiles for each model were considered. This approach narrowed the selection to 974 new terms, which then underwent expert evaluation[4]. This process validated 757 digital terms as "correct matches," yielding a 78% Positive Predictive Value (PPV) when integrating the outputs from multiple LLMs. Our methodology significantly reduced the manual effort required, filtering down from an initial dataset of over 40,000 digital terms to fewer than 1,000, thereby facilitating a focused and efficient expert review. The high rate of correct matches achieved underscores the effectiveness of VESPATE in keeping taxonomies both accurate and up-to-date with minimal human intervention. This successful deployment demonstrates its potential for enabling more dynamic and responsive labor market analysis.

[4]Two experts independently evaluated the entire set of matched terms: for Q1 the experts agreed on the evaluation of 86% of matches, for Q2 the agreement was 81%. A subsequent discussion was conducted for the terms where their evaluations initially diverged, leading to a consensus on the final evaluation.

5. Conclusion and Next Steps

This study introduces VESPATE (**VE**ctor **SPA**ce model for **T**axonomy **E**nrichment), a method designed to enhance and automate the updating of hierarchical taxonomies using contextual word embeddings. Hierarchical taxonomies are crucial for organizing complex concepts across domains like healthcare, finance, and labor markets. They must be regularly updated to reflect new developments, especially in fast-evolving fields like digital skills. Manual updates are often slow and error-prone, highlighting the need for automated solutions like VESPATE, which uses large language models (LLMs) to efficiently integrate new terms while preserving the taxonomy's structure. The traditional process of manually updating taxonomies is labor-intensive, time-consuming, and prone to errors, making the development of automated methods essential. VESPATE addresses this challenge by leveraging large language models (LLMs) as encoders to identify and integrate new, relevant terms into existing taxonomies without altering their foundational structure. Applied to the ESCO European Skill taxonomy, VESPATE successfully matched approximately 700 new digital terms, which were subsequently validated by domain experts. The process achieved an 81% Positive Predictive Value (PPV), demonstrating its effectiveness in aligning taxonomies with the rapidly changing demands of the labor market. This ability to keep taxonomies up to date is crucial for accurately tracking labor market shifts and informing policymakers' strategies. Looking ahead, future developments could explore alternative approaches for pre-cleaning data to enhance its quality for inclusion in the process. For instance, methodologies to identify and select terms that are more semantically relevant to a taxonomy could help reduce the size of large datasets and enable analysis on cleaner, more focused data. Additionally, leveraging a model specifically trained on labor market data, such as job postings and CVs, could offer a pathway to creating more contextually relevant skill representations and further improving performance. Currently, VESPATE has been successfully deployed as part of the EU project to enrich the ESCO digital skills taxonomy with emerging terms, ensuring stronger alignment with labor market needs. For future work, we plan to extend this method to tackle the task of taxonomy alignment and evaluate its effectiveness compared to state-of-the-art algorithms.

References

[1] K. Takeoka, K. Akimoto, M. Oyamada, Low-resource taxonomy enrichment with pretrained language models, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 2747–2758. URL: https://aclanthology.org/2021.emnlp-main.217. doi:DOI: 10.18653/v1/2021.emnlp-main.217.

[2] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso, Neo: A tool for taxonomy enrichment with new emerging occupations, in: The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2020, p. 568–584. URL: https://doi.org/10.1007/978-3-030-62466-8_35. doi:DOI: 10.1007/978-3-030-62466-8_35.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, 2020. arXiv:1906.08237.

[5] N. Vedula, P. Nicholson, D. Ajwani, S. Dutta, A. Sala, S. Parthasarathy, Enriching taxonomies with functional domain knowledge, 2018, pp. 745–754. doi:DOI: 10.1145/3209978.3210000.

[6] E. Manzoor, R. Li, D. Shrouty, J. Leskovec, Expanding taxonomies with implicit edge semantics, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2044–2054. URL: https://doi.org/10.1145/3366423.3380271. doi:DOI: 10.1145/3366423.3380271.

[7] R. Aly, S. Acharya, A. Ossa, A. Köhn, C. Biemann, A. Panchenko, Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings, 2019. arXiv:1906.02002.

[8] J. Shen, Z. Shen, C. Xiong, C. Wang, K. Wang, J. Han, Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network, in: Proceedings of The Web Conference 2020, ACM, 2020. URL: http://dx.doi.org/10.1145/3366423.3380132. doi:DOI: 10.1145/3366423.3380132.

[9] Y. Yu, Y. Li, J. Shen, H. Feng, J. Sun, C. Zhang, Steam: Self-supervised taxonomy expansion with mini-paths, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020). URL: https://api.semanticscholar.org/CorpusID:219792304.

[10] X. Song, J. Shen, J. Zhang, J. Han, Who should go first? a self-supervised concept sorting model for improving taxonomy expansion, ArXiv abs/2104.03682 (2021). URL: https://api.semanticscholar.org/CorpusID:233181932.

[11] J. Zhang, X. Song, Y. Zeng, J. Chen, J. Shen, Y. Mao, L. Li, Taxonomy completion via triplet matching network, in: AAAI Conference on Artificial Intelligence, 2021. URL: https://api.semanticscholar.org/CorpusID:230770060.

[12] S. Cheng, Z. Gu, B. Liu, R. Xie, W. Wu, Y. Xiao, Learning what you need from what you did: Product taxonomy expansion with user behaviors supervision, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), 2022, pp. 3280–3293. doi:DOI: 10.1109/ICDE53745.2022.00310.

[13] A. Alabdulkareem, M. R. Frank, L. Sun, B. AlShebli, C. Hidalgo, I. Rahwan, Unpacking the polarization of workplace skills, Science Advances 4 (2018) eaao6030. URL: https://www.science.org/doi/abs/10.1126/sciadv.aao6030. doi:DOI: 10.1126/sciadv.aao6030. arXiv:https://www.science.org/-doi/pdf/10.1126/sciadv.aao6030.

[14] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, Weta: Automatic taxonomy alignment via word embeddings, Computers in Industry 138 (2022) 103626. doi:DOI: 10.1016/j.compind.2022.103626.

[15] L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, Taxoref: Embeddings evaluation for ai-driven taxonomy refinement, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2021, p. 612–627. URL: https://doi.org/10.1007/978-3-030-86523-8_37. doi:DOI: 10.1007/978-3-030-86523-8_37.

[16] M. Arslan, C. Cruz, Semantic taxonomy enrichment to improve business text classification for dynamic environments, in: 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2022, pp. 1–6. doi:DOI: 10.1109/INISTA55318.2022.9894173.

[17] Y. Mao, T. Zhao, A. Kan, C. Zhang, X. Dong, C. Faloutsos, J. Han, Octet: Online catalog taxonomy enrichment with self-supervision, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020). URL: https://api.semanticscholar.org/CorpusID:219792569.

[18] J. Lu, Z. Pang, M. Xiao, Y. Zhu, R. Xia, J. Zhang, Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models, arXiv preprint arXiv:2407.06089 (2024).

[19] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, J. Zou, Mixture-of-agents enhances large language model capabilities, arXiv preprint arXiv:2406.04692 (2024).

[20] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, Mteb: Massive text embedding benchmark, arXiv preprint arXiv:2210.07316 (2022). URL: https://arxiv.org/abs/2210.07316. doi:DOI: 10.48550/ARXIV.2210.07316.