

Semantic Matching of IT Vocational Training Offers with the ESCO Ontology

Katerina Kostadinovska*[†] Kristine Hein*

0009-0002-0258-0857 0009-0005-6414-4294

*Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany

[†]TH Köln – University of Applied Sciences, Cologne, Germany

Email: katerina.kostadinovska@bibb.de, kristine.hein@bibb.de

Abstract—This paper presents a pipeline for semantically matching IT vocational training offers from the KURSNet database to skills defined in the ESCO ontology. The proposed approach combines semantic IT filtering, LLM-based keyword cleaning, and embedding-based matching using Sentence Transformer models. Applied to a dataset of 6,128 IT training courses with 16,372 unique keywords, the pipeline successfully maps 63.4% of keywords to ESCO skills at a cosine similarity threshold of 0.75. Evaluation against a manually annotated gold standard of 1,000 keyword-ESCO pairs yields a precision of 61.8%, with an inter-annotator agreement of $\kappa = 0.91$. The results demonstrate that embedding-based methods are suitable for bridging semi-structured national training data with standardized European competency frameworks.

I. INTRODUCTION

THE alignment of national vocational training data with standardized European competency frameworks represents a key challenge for labor market transparency and qualification comparability [1]. In Germany, KURSNet – operated by the Federal Employment Agency – serves as the central database for continuing education offers, describing courses through heterogeneous, non-standardized keywords [2]. The European Skills, Competences, Qualifications and Occupations ontology (ESCO) provides a multilingual reference framework of over 13,000 skill concepts, enabling structured competency modeling across European labor markets [3].

Despite the complementary nature of these systems, no systematic mapping between KURSNet keywords and ESCO skills exists. Embedding-based approaches using Sentence Transformers have shown promising results for semantic similarity tasks across domains [4], [5], suggesting their potential for bridging this gap. This paper presents a multi-step pipeline that combines semantic filtering, LLM-based keyword cleaning, and bi-encoder matching to automatically align IT vocational training offers with ESCO skills.

II. BACKGROUND AND RELATED WORK

A. KURSNet and ESCO

KURSNet, operated by the German Federal Employment Agency (BA), serves as the central database for continuing education offers in Germany. Training courses are described through semi-structured fields including titles, content descriptions, and freely assigned keywords [2]. These keywords represent the primary semantic units for content description,

yet exhibit significant terminological heterogeneity – varying in granularity, language (German/English), and specificity. The European Skills, Competences, Qualifications and Occupations (ESCO) ontology provides a multilingual reference framework of over 13,000 skill concepts, organized in a three-pillar structure covering occupations, skills, and qualifications [3]. ESCO follows Linked Open Data principles, using RDF and SKOS to model hierarchical (*broader/narrower*) and associative (*related*) relations between concepts [6], [7]. Each concept is identified by a unique URI, enabling interoperability across systems. The structural gap between KURSNet’s free-text keywords and ESCO’s formal ontology motivates the need for semantic matching approaches.

B. Structural Gap Between KURSNet and ESCO

Although both systems address vocational content, they differ fundamentally in their structural, terminological, and semantic design, as summarized in Table I. KURSNet stores information in field-structured format with hybrid keyword assignment – combining systematic classification, machine-generated terms, and manual additions by training providers [2] – resulting in terminological heterogeneity and varying levels of granularity. ESCO, by contrast, models occupations, skills, and qualifications as nodes in a formal semantic network, connected through hierarchical (*broader/narrower*) and associative relations [3].

TABLE I: Structural comparison of KURSNet and ESCO

Dimension	KURSNet	ESCO
System type	Operational information system	Ontological knowledge structure
Data structure	Field-based, semi-structured keywords	Graph-based with explicit relations
Terminology	Hybrid, partly free-form	Controlled multilingual vocabulary
Relations	No explicit semantic relations	broader/narrower, essential/optional skills
Interoperability	Limited	High (RDF, SKOS, URI)

This structural gap motivates the need for semantic matching approaches that capture meaning similarity beyond exact lexical overlap [8], [9]. For example, the KURSNet keyword *Teamführung* (team leadership) would not match the ESCO

skill *Mitarbeiter führen* (manage employees) under a purely lexical approach, despite their semantic equivalence. Similarly, *Machine Learning* would fail to match *maschinelles Lernen einsetzen* (apply machine learning) due to cross-lingual variation. These examples illustrate that lexical methods are insufficient for bridging the terminological gap between KURSNet and ESCO keywords.

C. Semantic Matching Approaches

Semantic matching of competency data has been approached through both ontology-based and embedding-based methods. Ontology-based approaches leverage explicit knowledge structures and hierarchical relations between skill concepts [10], [11], while Fernández-Sanz et al. demonstrate how different competency frameworks can be systematically mapped to the ESCO taxonomy [12]. Snijder et al. investigate ontology alignment methods between labor market ontologies such as ESCO and O*NET [13].

Embedding-based approaches model semantic similarity through learned vector representations [8], [14]. SentenceBERT [4] enables efficient similarity computation between text segments, while domain-adapted models such as VacancyBERT [15] demonstrate improved matching quality for job market data. Bhola et al. show that embedding-based methods outperform lexical approaches for skill extraction from unstructured text [16]. However, the systematic mapping of national vocational training databases with semi-structured keyword data to ESCO remains underexplored, motivating the present work.

III. METHODOLOGY

The proposed pipeline consists of four sequential steps for semantically aligning IT vocational training keywords from KURSNet with ESCO skills, as illustrated in Fig. 1.

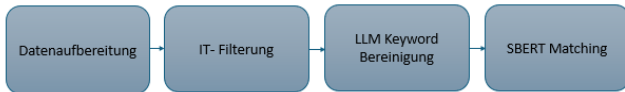


Fig. 1: Multi-step pipeline for semantic matching of IT vocational training offers to ESCO skills.

A. Data Preprocessing and Deduplication

The raw KURSNet dataset contains 60,540 training offers described through freely assigned keywords. As multiple offers may share identical keyword sets, deduplication is performed based on keyword set identity – two offers are considered identical if they share exactly the same combination of keywords, regardless of order. This keyword-based deduplication is justified by the analytical focus of this work: since the subsequent matching pipeline operates exclusively on keywords, two offers with identical keyword sets are semantically indistinguishable for the purpose of this analysis. This reduces the dataset to 26,872 distinct offers. The deduplicated data is

subsequently transformed into long format, where each keyword constitutes an individual observation, yielding 851,626 rows representing 98,841 unique keywords. This format serves as the basis for all subsequent steps.

B. Semantic IT Filtering

Since KURSNet covers all domains of continuing education, a semantic filtering step is applied to identify IT-relevant offers. The multilingual Sentence Transformer *paraphrase-multilingual-MiniLM-L12-v2* [4] is used to encode each course title against ten manually defined IT-domain anchor texts (e.g., *Software Development*, *Cybersecurity*, *Cloud Computing*). Each course is assigned the maximum cosine similarity score across all anchors. The score distribution shows a mean of $\bar{x} = 0.373$ and a 75th percentile of $P_{75} = 0.468$. A threshold of $\tau = 0.48$ is selected through iterative inspection of borderline cases – a qualitative review of ten boundary cases confirmed correct IT classification in eight instances – yielding 6,128 IT-relevant courses with 16,372 unique keywords.

C. LLM-Based Keyword Cleaning

The raw keywords exhibit significant heterogeneity: generic meta-terms (e.g., *Weiterbildung*, *Lehrgang*, *Fachkraft*), redundant entries, synonyms, and varying levels of specificity coexist within the dataset. Direct use of these raw keywords for ESCO matching would be problematic, as irrelevant terms dilute the semantic signal of the text representations.

Rule-based approaches such as stopwords lists or heuristic filters were considered but deemed insufficient, as they cannot capture semantic relationships between terms and perform poorly under terminological variance or ambiguity. Statistical methods such as clustering or frequency analysis reduce data volume but fail to account for the specific meaning context of individual courses [17], [18].

GPT-4o-mini was selected for its balance between model performance and processing efficiency. For the cleaning task – which focuses on selection, normalization, and semantic condensation – high-complexity reasoning is not required, making smaller models sufficient. Compared to larger language models, GPT-4o-mini enables substantially more cost-efficient and faster processing, which is particularly relevant when scaling to thousands of training offers.

A structured prompt provides the model with the course title as semantic context alongside the full list of original keywords. The course title serves as an interpretive frame, preventing keywords from being evaluated in isolation without reference to the actual course topic. The model is instructed to remove generic or organizational terms such as *Kurs* (course) or *Lehrgang* (training program), merge terminological variants and synonyms to reduce redundancy, prefer precise technical terms or short technical phrases over generic designations, and extract semantically relevant core phrases from longer keyword strings – for instance, reducing *Einführung in die digitale Fertigung* (Introduction to digital manufacturing) to *digitale Fertigung* (digital manufacturing). Output is constrained ex-

clusively to JSON format, enabling automated downstream processing.

A negative list of generic terms provides additional post-processing filtering, retry logic handles API failures automatically, and checkpoint storage ensures fault-tolerant processing of large batches. This step intentionally introduces a methodological trade-off: while synonym merging reduces redundancy, it may also conflate terms that represent distinct ESCO skills – for example, merging *SPS Technik* and *Speicherprogrammierbare Steuerungen* into a single representation. The LLM-based cleaning therefore represents a deliberate compromise between redundancy reduction and potential information loss.

D. Bi-Encoder Semantic Matching

The cleaned keywords are matched against ESCO skills using the multilingual Sentence Transformer *paraphrase-multilingual-mpnet-base-v2* [5], which produces 768-dimensional embeddings optimized for semantic similarity tasks. This model was selected over the MiniLM variant used in the filtering step, as the higher dimensionality enables finer-grained semantic differentiation – particularly relevant for distinguishing closely related ESCO skill concepts such as *Softwareentwicklung* and *Anwendungsentwicklung*. Embeddings are computed separately for all unique keywords and for all ESCO preferred labels (*preferredLabel*). Cosine similarity is computed via dot product on L2-normalized vectors. Only the preferred label is used for ESCO representation, as KURSNet keywords are typically short phrases, making label-level comparison structurally appropriate. Matching is performed on unique keywords to avoid redundant computation; results are subsequently mapped back to all corresponding courses.

The threshold $\tau = 0.75$ was selected based on a systematic evaluation across values from 0.65 to 0.85, as summarized in Table II. Lower thresholds yield higher coverage but lower average similarity, indicating a higher proportion of weaker matches. At $\tau = 0.70$, coverage increases to 12,687 keywords but mean similarity drops to 0.79. At $\tau = 0.80$, coverage drops substantially to 7,712 keywords, excluding a significant portion of relevant IT keywords. The value $\tau = 0.75$ represents the optimal trade-off between coverage (63.4%) and semantic quality ($\bar{x} = 0.82$). For each keyword, up to $K = 10$ ESCO skills with the highest similarity scores are retained (Top-K).

TABLE II: Threshold evaluation results

Threshold	Keywords matched	$\bar{x}$ Similarity
0.65	14,284	0.76
0.70	12,687	0.79
0.75	10,388	0.82
0.80	7,712	0.85
0.85	4,992	0.89

IV. RESULTS AND EVALUATION

A. Matching Results

The pipeline was applied to the filtered dataset of 6,128 IT-relevant training courses. At threshold $\tau = 0.75$, 10,388 out of 16,372 unique keywords (63.4%) were successfully matched to at least one ESCO skill, generating 54,390 unique keyword-ESCO pairs across 5,109 distinct ESCO skills. The mean cosine similarity of matched pairs is $\bar{x} = 0.819$ (median = 0.808), indicating that the majority of matches lie well above the threshold. A rank analysis shows a shallow decline from $\varnothing 0.85$ (rank 1) to $\varnothing 0.79$ (rank 10), suggesting that lower-ranked matches remain semantically plausible and justify the Top-K approach.

The most frequently matched ESCO skills reflect the thematic focus of the IT training landscape, with Cloud technologies, project management, and data analysis dominating the top assignments. A total of 213 keyword-ESCO pairs achieve a cosine similarity of ≥ 0.99 , typically arising from lexically identical or near-identical label pairs such as *DevOps* $\rightarrow$ *DevOps* or *agiles Projektmanagement* $\rightarrow$ *agiles Projektmanagement*.

Table III illustrates the Top-K matching results for the keyword *Data Science*, demonstrating how a single keyword can be associated with multiple semantically related ESCO skills.

TABLE III: Top-K ESCO matches for keyword *Data Science*

ESCO Skill	Cosine Similarity
Datenanalyse	0.851
Datenanalyse durchführen	0.808
Daten sammeln	0.798
Datenmodelle	0.784
Datensätze erstellen	0.777
Forschungsdaten verwalten	0.760
Daten untersuchen	0.759
Daten verarbeiten	0.758

Qualitative inspection reveals systematic error patterns: morphologically similar but semantically distinct pairs such as *Hafenordnung* (harbor regulations) $\rightarrow$ *Ports* (network ports) or *Auftragsbearbeitung* (order processing) $\rightarrow$ *Brauverfahren* (brewing process) demonstrate the limitations of purely similarity-based approaches without additional conceptual modeling.

B. Evaluation

To assess matching quality, a gold standard of 1,000 keyword-ESCO pairs was manually annotated by three independent annotators using a three-point scale: *yes* (relevant), *approximately* (partially relevant), and *no* (not relevant). Both *yes* and *approximately* were treated as positive matches. The dataset was stratified across five cosine similarity intervals (250/250/250/162/88 pairs), ensuring representative coverage of the full similarity range.

Inter-annotator agreement was measured using Cohen's Kappa [19], yielding $\kappa = 0.91$, indicating almost perfect agreement [20]. The overall observed agreement rate was

95.6%, with 44 disagreements resolved by a third annotator whose judgment served as the final gold standard label.

Overall precision of the matching procedure is **61.8%** (618 out of 1,000 pairs rated as relevant). As shown in Table IV, precision increases substantially with cosine similarity score, demonstrating a clear positive correlation between similarity score and match quality.

TABLE IV: Precision by cosine similarity interval

Interval	Precision	n
[0.75, 0.80)	39.6%	250
[0.80, 0.85)	57.2%	250
[0.85, 0.90)	63.2%	250
[0.90, 0.95)	82.1%	162
[0.95, 1.00]	96.6%	88
Overall	61.8%	1,000

Matches in the lowest interval [0.75, 0.80) are frequently thematically related but not semantically equivalent, while matches above 0.90 are predominantly correct. The highest interval approaches near-perfect precision, driven by lexically identical label pairs.

V. DISCUSSION AND LIMITATIONS

A. Discussion

The proposed pipeline achieves a precision of 61.8%, which represents a solid baseline given the challenging input data. Unlike existing work focusing on structured competency profiles or job postings [15], [16], [10], this work addresses semi-structured, terminologically heterogeneous keyword data – a significantly more demanding setting.

The score-dependent precision analysis confirms that cosine similarity is a reliable indicator of match quality. A clear trade-off is visible: higher thresholds yield substantially better precision but reduce coverage. For practical applications, threshold selection depends heavily on whether precision or recall is prioritized.

The thematic concentration of matches on Cloud computing, IT security, and project management reflects both actual trends in the IT labor market [21] and data-driven effects from keyword frequency distributions. Results should therefore be interpreted as reflecting both content trends and dataset-specific characteristics.

Error analysis reveals that the method’s limitations lie primarily at the boundary between conceptual proximity and semantic equivalence. Morphological similarity – such as *Hafenordnung* (harbor regulations) → *Ports* (network ports) or *Schraubenschlüssel verwenden* (use wrench) → *Primärschlüssel* (primary key) – represents a systematic error source that purely similarity-based approaches cannot reliably resolve, underscoring the need for complementary methods incorporating contextual or ontological information.

B. Limitations

Several limitations should be acknowledged. First, the LLM-based keyword cleaning was not systematically evaluated; relevant content may be lost or distorted, and the

non-deterministic nature of LLMs limits reproducibility. The chosen prompt formulation further influences results, as synonym merging may conflate terms that represent distinct ESCO skills. Second, the high terminological heterogeneity and varying granularity levels of KURSNet keywords directly affect embedding quality and constrain achievable matching performance – similar concepts are frequently described using different formulations (e.g., *Webentwicklung* vs. *Frontend*), leading to inconsistent matching results. Third, the bi-encoder approach does not model fine-grained interactions between keywords and ESCO skills; cross-encoder models [22] could improve precision for borderline cases by jointly processing both texts, though at substantially higher computational cost for large-scale matching tasks. Fourth, matching relies exclusively on short label texts, disregarding contextual information from course descriptions or metadata; this may cause semantic nuances to be lost, particularly for polysemantic or highly generic terms. Fifth, the dataset is restricted to IT-relevant offers, limiting generalizability to other vocational domains such as healthcare or skilled trades. Finally, evaluation is based solely on precision; recall cannot be computed as the gold standard covers only model-proposed pairs, precluding assessment of false negatives and thus providing only a partial quality evaluation of the matching procedure.

VI. CONCLUSION AND OUTLOOK

This paper presented a multi-step pipeline for semantically matching IT vocational training keywords from the KURSNet database to skills defined in the ESCO ontology. The approach combines semantic IT filtering, LLM-based keyword cleaning, and bi-encoder matching using multilingual Sentence Transformers.

The results confirm that embedding-based methods are suitable for bridging semi-structured national training data with standardized European competency frameworks, even when matching is performed exclusively on short label texts. From a practical perspective, the approach offers concrete benefits for multiple stakeholders: research institutions such as BIBB can leverage the mappings for data-driven analysis of competency trends within the German continuing education landscape – for instance, identifying which ESCO skills are underrepresented in current IT training offers, or tracking shifts in demand for Cloud computing and cybersecurity competencies over time; public employment agencies could integrate the method into digital training portals to enable competency-based search; and training providers may benefit from automated, standardized competency tagging of their offers. On a policy level, the approach supports data-driven assessment of the alignment between national training landscapes and European competency frameworks, contributing to labor market transparency and qualification comparability.

Several directions for future work emerge. Incorporating full ESCO skill descriptions and alternative labels (*non-preferred terms*) could increase semantic expressiveness, particularly for generic or ambiguous keywords. Extracting competencies

directly from course descriptions rather than assigned keywords may capture implicit skills not reflected in explicit keyword fields, with context-based representations potentially improving matching quality for polysemantic terms. Domain-specific fine-tuning of embedding models on labor market and education-specific texts [15] and cross-encoder re-ranking [22] offer promising avenues for improving precision for borderline cases. A more systematic, data-driven optimization of the similarity threshold τ – explicitly balancing the trade-off between precision and coverage – represents another valuable direction. Extending the pipeline to non-IT domains would assess its generalizability and enable broader evaluation of ESCO’s semantic coverage. Finally, a cluster analysis of derived ESCO assignments could reveal latent competency structures, thematic groupings, and potential occupational transition pathways within the continuing education landscape, representing the next step toward a comprehensive data-driven competency analysis of national vocational training systems.

ACKNOWLEDGMENTS

During the preparation of this manuscript, the authors used DeepL to check the language and grammar. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

REFERENCES

- [1] OECD, *Getting Skills Right: Future-Ready Adult Learning Systems*. OECD Publishing, 2019.
- [2] Bundesagentur für Arbeit, “Eingabehilfe für die erfassung von weiterbildungsangeboten in kursnet,” Bundesagentur für Arbeit, Nürnberg, Tech. Rep., 2025, stand: 31.03.2025.
- [3] European Commission, *ESCO Handbook: European Skills, Competences, Qualifications and Occupations*, 2nd ed. Luxembourg: Publications Office of the European Union, 2019.
- [4] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP*, 2019.
- [5] —, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. [Online]. Available: <https://arxiv.org/abs/2004.09813>
- [6] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data – the story so far,” *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.
- [7] World Wide Web Consortium (W3C), “Skos simple knowledge organization system reference,” <https://www.w3.org/TR/skos-reference/>, 2009, w3C Recommendation.
- [8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *ICLR*, 2013.
- [9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.
- [10] T. Petrican and O. Stan, “Ontology-based skill matching algorithms,” in *Proceedings of CSCS*, 2015.
- [11] S. Paudel and S. Shakyia, “Ontology based job-candidate matching using skill sets,” in *IOE Graduate Conference*, 2017.
- [12] L. Fernández-Sanz *et al.*, “e-skills match: A framework for mapping skills into esco,” *Computer Standards & Interfaces*, vol. 51, pp. 30–42, 2017.
- [13] L. Snijder *et al.*, “Advancing ontology alignment in the labor market,” in *AAAI-MAKE 2024*, 2024.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019.
- [15] N. Bocharova *et al.*, “Vacancybert: A domain-adapted language model for job matching,” *Applied Sciences*, 2021.
- [16] A. Bhola *et al.*, “Extracting skills from job descriptions using embeddings,” in *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 2020.
- [17] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: <https://arxiv.org/abs/2005.14165>
- [18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [19] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [20] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [21] OECD, *OECD Skills Outlook 2021: Learning for Life*. Paris: OECD Publishing, 2021.
- [22] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. [Online]. Available: <https://arxiv.org/abs/2104.08663>