Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads

Ann-Sophie Gnehm and **Eva Bühlmann** and **Helen Buchs** and **Simon Clematide**
Department of Sociology and Department of Computational Linguistics
University of Zurich
{gnehm,buehlmann,buchs}@soziologie.uzh.ch, simon.clematide@cl.uzh.ch

Abstract

Monitoring the development of labor market skill requirements is an information need that is more and more approached by applying text mining methods to job advertisement data. We present an approach for fine-grained extraction and classification of skill requirements from German-speaking job advertisements. We adapt pre-trained transformer-based language models to the domain and task of computing meaningful representations of sentences or spans. By using context from job advertisements and the large ESCO domain ontology we improve our similarity-based unsupervised multi-label classification results. Our best model achieves a mean average precision of 0.969 on the skill class level.

1 Introduction

How skill demand evolves over time in the labor market has always been a main research question in social sciences. Research has however been hampered by the following limitations: Skills were mostly measured on the supply side (what workers bring, not what employers ask for) and only on an aggregated level (by occupations) and/or cross-sectional (one data point in time). Furthermore, most data focused on a selection of skills, since defining and measuring skills is difficult (Biagi and Sebastian, 2020). Job advertisement data can help to overcome such shortcomings by providing time-series measurements on the job level, including all labor market skill requirements (Buchmann et al., 2022a). Not surprisingly, social science has thus lately shown great interest in applying text mining methods to job advertisements (job ads in the following).

Our main goals are to, first, extract spans of text in Swiss German-speaking job ads that specify workers' skill requirements: Specifically, educational requirements, work experiences or skills, and language competences. Second, we classify the extracted spans onto the large, fine-grained *European Skills, Competences, Qualifications and Occupations Ontology* (ESCO).[1] Third, we show the value of the data-driven extraction results in evaluations and initial social science analyses.

Our general idea is to use, in an unsupervised approach, the semantic similarity between ontological concepts and text spans in job ads for fine-grained classification of job ad skills to the ESCO skill ontology. We rely on state-of-the-art pre-trained transformer-based language models (foundational models) and experiment with adaptations to the job ad domain and to the task of computing the semantic similarity on sentence or span level. Additionally, we assess different methods to exploit the textual content and terminological richness of the ESCO ontology for fine-tuning the foundational language models. And, we show how providing additional textual context from the job ads and/or the ontology improves the similarity scores between skill requirement spans in job ads and their corresponding concepts from the ontology.

Our contributions include a definition of skill requirement mention types and annotation guidelines for fine-grained extraction, and an exploration of NLP methods for improving semantic similarity measures for matching job ad text snippets with ESCO terminology. We contribute further sentence-level language representation models that are adapted to the job ad domain and skill-related expressions, and we incorporate terminological variability from a large ontology into the model.

Section 2 discusses related work. Section 3 describes our data. Our approaches, experiments, and results for extracting skills are explained in Section 4, and for classification in Section 5. Section 6 shows initial sociological analyses on the extracted data. Section 7 summarizes our main findings and directions for future work.

[1] See https://esco.ec.europa.eu/en/about-esco/what-esco, (European Commission. Directorate General for Employment, Social Affairs and Inclusion., 2017)

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 14 - 24
November 7, 2022 ©2022 Association for Computational Linguistics

2 Related Work

2.1 Skill Extraction from Job Ads

For the US and UK job market, recent studies investigate changing skill requirements in jobs ads (Deming and Kahn, 2018; Hershbein and Kahn, 2018; Azar et al., 2018), with newer research pointing out the importance of new skills entering jobs and altering required skill combinations within professions (Acemoglu et al., 2022; Atalay et al., 2020). However, these approaches use mostly proprietary data, where extraction is not fully documented. Recently, Zhang et al. (2022b) worked on fine-grained skill classification using their English and Danish *Kompetencer* dataset. They use the ESCO API to retrieve 100 candidates per manually annotated skill span and select the best candidate for their silver standard annotation by minimal Levenshtein distance. Fine-tuning a multilingual BERT-style model on their small in-domain and in-language training material resulted in big improvements compared to their few-shot setup.

2.2 NLP Methods for Improving Semantic Similarity Measures

Continued in-domain pre-training: Masked language modeling (MLM) on domain or task-specific data is often and successfully applied for adapting general-domain language models to specific domains or even tasks (see Gururangan et al. (2020) for an overview, or Gnehm et al. (2022) and Zhang et al. (2022a) for applications on job ads.)

Sentence-level fine-tuning: Reimers and Gurevych (2019) were the first to adapt pre-trained transformer-based language models with supervised training on natural language inference (NLI) and semantic textual similarity (STS) datasets. Resulting Sentence-BERT (SBERT) models can be used to efficiently compare semantic similarities on the sentence level. Many subsequent approaches leverage more self-supervised training to lower data requirements, often by using unlabeled data and by synthetically creating pairs of similar sentences from a single source sentence (Giorgi et al., 2021; Gao et al., 2021; Wang et al., 2021). Differences between the approaches in architectures and training objectives are discussed in Section 5.2.

3 Experimental Data

3.1 Job Ad Data

We use the Swiss Job Market Monitor (SJMM) dataset consisting of representative yearly samples of print and online job ads from Switzerland from 1950 up to now.[2] Being representative and longitudinal, the data is ideal for research on the evolution of skill requirements. In our experiments, we focus on German-speaking job ads from 1990-2021 (n=53k).

3.2 Ontological and Terminological Data

ESCO: We use the German data of the multilingual ESCO ontology (v1.1.0), comprising 14.5k skill concepts. Each concept is represented by a **preferred term** (e.g., *use spreadsheets software*), often complemented by **alternative terms** (synonyms as *use spreadsheets programs*) or **hidden terms** (outdated terms or specific products, *Microsoft Office Excel*).

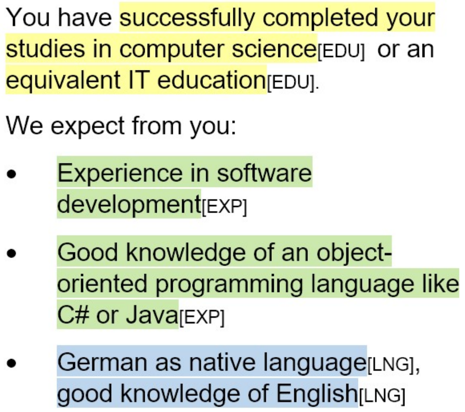
In total, the 14.5k ESCO concepts are expressed by 20k terms, and include **knowledge** (e.g., *pharmacotherapy*), **skills** in a narrower sense (an ability as *apply change management*), **language skills** (*understand spoken French*), and **transversal skills**, also referred to as core or soft skills (*negotiate compromises*). These four fields are hierarchically structured into 638 classes (max. depth of 3 with 475 classes on the lowest level). The concepts are internally ordered by broader/narrower relationships and are linked to these classes directly or via broader concepts. ESCO is multi-hierarchical and a concept may have several broader concepts (e.g., *aviation meteorology* belongs to the broader concepts *meteorology* and *transport services*). Overall, 29.3% of concepts (30.5% of terms) fall into more than one class.

Swiss databases:[3] We dispose of Swiss terminology on professions and qualifications that has been linked to ESCO **knowledge** classes (e.g., the term *architect* belongs to the ESCO class *architecture and town planning*). This adds 39k terms (20.5k concepts) to 102 knowledge classes and should help identify Swiss educational requirements. Here the class ambiguity is much lower, only 0.1% of concepts (0.4% of terms) belong to more than one class.

Custom terminology additions: We add a handful of terms to cover a few Swiss-specific high-

[2] See https://www.swissubase.ch (Buchmann et al., 2022b)

[3] Swiss Federal Statistical Office, data available on request



You have successfully completed your studies in computer science[EDU] or an equivalent IT education[EDU].

We expect from you:

- Experience in software development[EXP]
- Good knowledge of an object-oriented programming language like C# or Java[EXP]
- German as native language[LNG], good knowledge of English[LNG]

Figure 1: Example extraction of EDU, EXP, LNG spans (examples translated from German to English)

	Precision	Recall	F-score
EXP	0.856	0.831	0.843
EDU	0.861	0.859	0.861
LNG	0.885	0.914	0.899

Table 1: Skill extraction results per skill span type on final test set (n=200 ads)

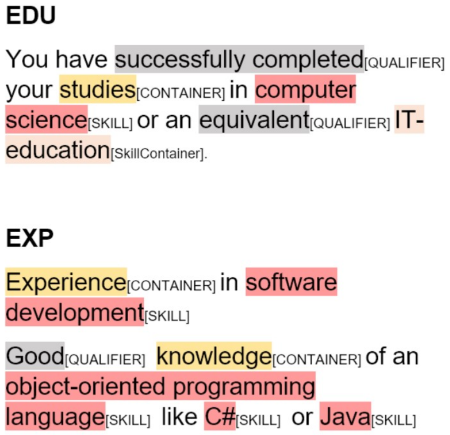
frequency abbreviations, which are not represented as such in the ontology, e.g., ‘KV’ for ‘kaufmännische/r Angestellte/r’ (*commercial clerk*).

4 Skill Extraction

4.1 Coarse Skill Span Extraction

We first trained a model to extract text spans from the ads that contain skill requirements. Three span types were defined for this coarser task: **education** (EDU), **experiences** (EXP), and **language skills** (LNG). EDU spans include requirements for both formal and informal education and further training. EXP spans contain all required experiences and knowledge, which are not specified in terms of specific education. LNG spans describe requirements for the language skills of the applicants. Figure 1 shows an annotated example.

We annotated 2,000 ads iteratively with the annotation tool *prodigy*[4]. To start, a domain expert annotated a sample of around 100 ads to refine the annotation guidelines and train an initial model. Then, we built the rest of the training data in 7 iterations, where the same annotator corrected each time roughly 250 ads pre-annotated by the model. We retrained the model after every iteration using 80% of available data as training, 10% as development, and 10% as test set.



EDU

You have successfully completed[QUALIFIER] your studies[CONTAINER] in computer science[SKILL] or an equivalent[QUALIFIER] IT-education[SkillContainer].

EXP

Experience[CONTAINER] in software development[SKILL]

Good[QUALIFIER] knowledge[CONTAINER] of an object-oriented programming language[SKILL] like C#[SKILL] or Java[SKILL]

Figure 2: Examples for fine-grained extraction of QUALIFIER, CONTAINER, and SKILL areas in EDU and EXP spans (examples translated from German to English)

	Precision	Recall	F-score
EXP			
QUALIFIER	0.953	0.968	0.960
SKILL	0.910	0.915	0.913
CONTAINER	0.940	0.973	0.956
EDU			
QUALIFIER	0.947	0.989	0.968
SKILL	0.940	0.951	0.945
CONTAINER	0.922	0.936	0.929
SkillContainer	0.874	0.908	0.891

Table 2: Fine-grained skill area extraction results on final test set (n=200 ads)

We treated the extraction and classification of skill spans as a named-entity-recognition-like problem and trained a transition-based NER model (Lample et al., 2016) using *spaCy*[5]. We used *jobBERT-de*[6], a German transformer model adapted to the domain of job ads (Gnehm et al., 2022), to compute contextualized input representations for the downstream NER component.

4.2 Fine-Grained Skill Area Extraction

Within the extracted EDU and EXP spans, different content aspects are present, as shown in Figure 1 and 2. In addition to information about the specific **skill area**, they also specify the **qualitative level** of a skill, or mention also generic **skill re-**

[4] https://prodi.gy

[5] https://spacy.io. We used the default settings of the components *spacy-transformers.TransformerModel.v1* and *spacy.TransitionBasedParser.v2*

[6] https://huggingface.co/agne/jobBERT-de

quirement containers. To better capture the core content of the skills, a more fine-grained skill area extraction model has been trained for both the EXP and the EDU spans. The training data was created the same way as for coarse-grained extraction (see Section 4.1). Formally, these models split the spans into different areas: QUALIFIER, SKILL, CONTAINER, and SkillContainer. The last category was introduced only in the EDU domain to capture compounds that contain both skill area and container information. In German, such compounds occur frequently, e.g., 'Handelsdiplom' (*commercial diploma*), 'Bürolehre' (*office apprenticeship*). Figure 2 shows how the EDU and EXP spans from Figure 1 are refined accordingly.

4.3 Results

Table 1 shows the results for the skill span extraction on the final test set (n=200 ads). For LNG, it performs best with an F-score of 0.899, while EXP performs least well with 0.843. This reflects the higher complexity of the EXP span task. Table 2 reports the performance of the fine-grained skill area extraction. In general, all categories perform very well, with F-scores above 0.9. Only the SkillContainer category scores slightly worse with 0.891.

5 Fine-Grained Unsupervised Multi-Label Classification of Skill Requirements

5.1 Task Definition

In order to map a skill mention of a job ad to one or more fitting ESCO concepts, we perform a semantic similarity lookup, comparable to an information retrieval setting, where, for a given query (job ad skill), we search for the most relevant items (ontology skill concepts). The problem can thus be understood as an unsupervised, fine-grained multi-label classification task.

Contextualizing job ad terms: As introduced in Section 4, we use skill areas for our query. However, isolated skill areas without surrounding job ad text can be too generic or ambiguous, potentially leading to unsuitable matches. To mitigate this issue, we contextualize each skill area with available surrounding skill areas of the same span.[7] After embedding these contextualized text spans with an SBERT model, we calculate a vector representation for each skill area by averaging the vector representation of each token. Contextualization helps us find more exact skill concepts, e.g., if we query *project management*, we receive *project management* as top suggestion, but if we query *project management* with its context *IPMA, PMI, HERMES*, we find the more specific concept *IT project management methods*. It helps further dealing with incomplete skill areas, as they occur for instance in elliptic enumerations: Querying *Motor vehicle* in its context *Motor vehicle, liability, property insurance* returns *insurance types* as top suggestion.[8]

Contextualizing ontology terms: In the lookup, we use all available ontology terms (see Section 3.2). As preprocessing, we remove information on educational levels in the Swiss data, such that – as for the job ads – only a skill area remains (e.g. *florist (Federal Professional Certificate)* is transformed to *florist)*. Ontology terms can also be ambiguous by themselves, and many belong to more than one skill class (see Section 3.2). Therefore, we contextualize ontology terms too, and use the hierarchical ontology structure by inserting its class label for each term as context. For embedding with SBERT models, we represent these term and class combinations in the form '<term> (<class label>)'.[9] For each term, a vector representation is calculated in the same way as described above for the job ad terms. To give an example, with contextualized ontology terms, querying the job ad skill *SAP developer ABAP*, we find that *SAP ABAP* in the class *Software and applications development* is more similar than *SAP ABAP* in the class *Using digital tools for collaboration and productivity*.

5.2 Semantic Skill Representation Approaches

The quality of the results of the vector similarity search depends crucially on a suitable vector space representation of the skill descriptions from the job ads and from the ontology. Therefore, we experiment with several state-of-the-art approaches for improving the vector similarity of general BERT language representation models by applying continued pretraining and fine-tuning techniques.

MLM on job ad texts: Masked language mod-

[7] In total, we have 131k areas from 78k EXP spans, and 81k areas from 74k EDU spans available. The 39k LNG spans were not further split up.

[8] ESCO queried with the model sts-gbert.

[9] After initial experiments, 173 knowledge class labels were replaced by custom labels using a language less formulaic and more common for job ads, e.g., *services in the field of transportation* was replaced by *transportation services*.

eling (Devlin et al., 2019) on in-domain texts has been successfully used for adaptation of general-domain BERT models to special domain language use (Gururangan et al., 2020). We assess the benefit of continued in-domain language model pretraining by comparing *GBERT-base*[10], a small version of the German state-of-the-art model (Chan et al., 2020), with a version of the same model that is adapted to the domain of German-speaking job ads, and was trained on a job ad dataset including the data used here, *jobGBERT*[11] (Gnehm et al., 2022).

TSDAE on skill spans: In the transformer-based sequential denoising auto-encoder (TSDAE) approach (Wang et al., 2021), meaningful sentence embeddings are learned by denoising corrupted input. An encoder produces a fixed-size vector representation for an input sentence with deleted words, from which a decoder learns to reconstruct the uncorrupted sentence. By giving the decoder only the fixed-size sentence representation and no word embeddings as input, a bottleneck is introduced that forces the encoder to provide a good semantic sentence representation. We use TSDAE to learn embeddings for our domain-specific skill terminology. As training data, we use all skill spans from our job ad data (216k), and skill terms and descriptions (split into sentences) from our ontology data (107k). Since our spans are shorter than the sentences used in the original approach (2.2 vs. 10.6 tokens on average), we experimented with smaller deletion rates and found a rate of 0.4 best performing. All other parameters are set as in Wang et al. (2021).

STS on general-domain data: Reimers and Gurevych (2019) use Siamese BERT Networks for training sentence embeddings on sentence pairs which are labeled with a cosine similarity score indicating their semantic similarity. Sentence vector representations are calculated by mean pooling over token embeddings. Then, by computing the similarity of the two sentence vectors and by comparing it against the gold similarity score, better semantic sentence representations are learned. No such labeled data is available for our domain, but we assess the benefits of fine-tuning our sentence embeddings on general-domain data for German by using the translated *STSBenchmark* dataset (May, 2021) (5k sentence pairs). We train with hyperparameters set as in Reimers and Gurevych (2019).

EXP skill: ***attracting new customers***, *acquisition*		
skill concept suggestions	A	B
recruitment methods (marketing and advertisement)	0.5	0.5
customer insight (marketing and advertisement)	0.5	0
find new clients (entrepreneurial skills)	1	1
recruitment and hiring (personnel recruitment)	0	0
EDU skill: ***bio lab technician***		
skill concept suggestions	A	B
biologist (biology)	0.5	0.5
biology technician (biology)	1	0.5
biology lab technician (chemical technology)	1	1
biology teaching assistant (specialist subject teachers)	0	0
LNG skill: ***English (very good in spoken and written)***		
skill concept suggestions	A	B
teach English as a foreign language (teaching)	0	0
understand written English (languages)	1	1
English speaking skills (languages)	1	1
English teacher (specialist subject teachers)	0	0

Table 3: Evaluation examples of skill concept suggestions (class labels in brackets) for an EDU, an EXP, and an LNG job ad skill (in bold italics, context in italics) by two annotators A and B (examples translated from German to English).

MNR on ontology data: Sentence embeddings are learned by training Siamese networks with multiple negative ranking (MNR) loss (Henderson et al., 2017). This is a supervised approach, but training data requirements are low since only pairs of similar sentences are needed. Dissimilar sentence pairs are created by using other examples from the same batch of training sentences. The relative distances between sentence pairs are then learned using a ranking loss function. We leverage our ontology data by creating positive text pairs in which we combine alternative or hidden terms, as well as the phrases describing them, each with their preferred label. In this way, we seek to incorporate knowledge of terminological variations within the ontology into sentence embeddings. We expect this approach to be the most beneficial since it is using data specific to our domain and task in a supervised fashion.

5.3 Experiments and Evaluation

Evaluation data: To be able to evaluate models on our fine-grained unsupervised multi-label classification task, we created a small amount of gold standard data. We selected a random sample of 25 job ad skill terms and in addition compiled a challenge sample of 15 terms covering some difficult cases (e.g., formulations that are specific to Switzerland). For these 40 terms, we did a contextualized ontology lookup as described in Section 5.1 using all our different SBERT models (see below), and evaluated the ten first suggestions of all models.

[10] https://huggingface.co/deepset/gbert-base

[11] https://huggingface.co/agne/jobGBERT

Annotators assigned scores of 0 for inadequate, 0.5 for acceptable, and 1 for highly appropriate suggestions. We evaluated the suggestions on the class level and for the random sample also on the concept level. Table 3 shows examples for evaluation on concept level. A total of 494 class suggestions were rated by 3 annotators each and 685 concept suggestions were rated by 2 annotators each. For the random sample, Krippendorff's alpha on the class level is 0.83, on concept level 0.814, and for the challenge set on class level 0.73. This indicates good agreement for the random and satisfactory agreement for the challenge sample (Krippendorff, 2004).

Experiments: We evaluate combinations of the presented SBERT training approaches with some restrictions: MLM on domain data only makes sense as the first step, since it affects the foundational model on the token level. STS after TSDAE is more effective than vice versa, according to Wang et al. (2021), and domain-oriented (MNR) is applied after general-domain (STS) fine-tuning.[12] This leads to a total of 14 tested model configurations: Starting from a general (gbert) or domain-adapted (jobgbert) LM, we optionally train with TSDAE (model name prefix: tsdae-), followed by optional STS (prefix: sts-), followed by optional MNR (prefix: mnr-). A model with only STS training on a general-domain LM (sts-gbert in the following) corresponds to a vanilla or baseline SBERT model. For selected models, we further perform an ablation study to estimate the effects of contextualizing job ad skills and/or ontology skills for similarity queries.

We use the created gold standard data to evaluate fine-grained unsupervised skill classification with mean average precision over the first ten concept or class suggestions (mAP@10), see Equation 1, where Q are the queries, (25 in our case for the random sample), m is the number of accepted suggestions, and k is the cutoff rank (10 in our case).[13] Mean average precision considers the ranking capabilities of models (are more appropriate suggestions presented first?) and does not unfairly penalize models when too few suitable items are available (less than ten items for mAP@10) (Manning et al., 2008).

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_{\text{j}}}\sum_{k=1}^{m_{\text{j}}} Precision(R_{\text{jk}}) \tag{1}$$

A conventional recall evaluation (are all relevant ontology concepts among the suggestions?) is not applicable in this scenario with 638 classes and 35k concepts. However, we examine mentions with very low similarities to ontology concepts.

5.4 Results and Discussion

Fine-grained skill classification performance:

In classification from job ads to ESCO, the best model on class level is mnr-sts-jobgbert with 0.969 mAP@10, on concept level mnr-sts-tsdae-jobgbert with 0.908 mAP@10 (see Table 4). As expected, evaluation scores are lower on concept than on class level, since it is much harder to find an appropriate concept out of 35k possibilities than an appropriate class out of 638. Performance differences between models are often small, but it is noticeable that the best models at both levels include MNR as pre-training. MNR seems thus to have a strong positive impact on performance, while the effect of other pre-training steps is less obvious, and including additional pre-training does not ensure higher performance compared to vanilla sts-gbert.

On the challenge test set (not shown in Table 4), all models experience a performance drop compared to the random sample, but to varying degrees. For instance, the sts-gbert model with general-domain pre-training only achieves mAP@10 of 0.763. Compared to the random sample this is a loss of 15.1 percentage points (pp in the following). Our best models mnr-sts-tsdae-jobgbert and mnr-sts-jobgbert reach both mAP@10 of 0.9, which means a smaller performance drop of 6.2 and 6.9pp respectively. Hence, extensive SBERT fine-tuning also pays off for classifying more difficult cases.

The mapping of EDU and LNG terms is, in general, easier than the mapping of EXP terms, with models reaching on average mAP@10 of 0.952 and 0.938 versus 0.878 (on class-level, see Table 4). Interestingly, model performance can vary considerably across different skill types, suggesting that fine-tuning approaches may have type-specific effects (see discussion below).

Impact of different SBERT fine-tuning steps: To assess different sentence embedding fine-tuning steps, we estimate their effects on mean average

[12] In MNR we used batch-size of 32 after pre-tests with batch sizes 16, 32, 64. If not specified differently in Section 5.2, all other parameters are set as in the original approaches.

[13] We considered a suggestion as true positive if at least one annotator gave a score of 1, or at least two annotators a score of 0.5.

Class Level								
Model	**ALL**	**R**	**EDU**	**R**	**EXP**	**R**	**LNG**	**R**
mnr-sts-jobgbert	**0.969**	**1**	0.977	5	**0.945**	**1**	**1.000**	**1**
mnr-sts-tsdae-jobgbert	0.961	2	0.977	5	0.944	2	0.963	9
mnr-gbert	0.958	3	0.987	2	0.929	3	0.957	10
mnr-tsdae-jobgbert	0.957	4	0.983	3	0.924	4	0.968	8
mnr-sts-tsdae-gbert	0.954	5	0.983	3	0.902	6	0.998	2
mnr-jobgbert	0.941	6	0.940	11	0.923	5	0.976	6
mnr-tsdae-gbert	0.940	7	0.967	8	0.890	9	0.988	5
sts-tsdae-gbert	0.935	8	**0.996**	**1**	0.856	10	0.970	7
sts-jobgbert	0.914	9	0.926	12	0.899	7	0.919	11
mnr-sts-gbert	0.903	10	0.865	14	0.893	8	0.998	2
tsdae-gbert	0.879	11	0.968	7	0.786	14	0.887	13
sts-gbert (baseline)	0.876	12	0.870	13	0.826	11	0.990	4
sts-tsdae-jobgbert	0.872	13	0.947	10	0.787	13	0.890	12
tsdae-jobgbert	0.821	14	0.948	9	0.790	12	0.631	14
mean	0.920		0.952		0.878		0.938	
stdev	0.044		0.041		0.059		0.096	

Concept Level								
Model	**ALL**	**R**	**EDU**	**R**	**EXP**	**R**	**LNG**	**R**
mnr-sts-tsdae-jobgbert	**0.908**	**1**	0.923	5	**0.865**	**1**	0.963	8
mnr-gbert	0.897	2	**0.950**	**1**	0.825	3	0.934	10
mnr-tsdae-jobgbert	0.889	3	0.947	2	0.791	6	0.968	7
mnr-sts-jobgbert	0.886	4	0.874	10	0.842	2	**1.000**	**1**
sts-tsdae-gbert	0.868	5	0.928	4	0.758	8	0.970	6
sts-gbert (baseline)	0.867	6	0.878	8	0.795	5	0.990	4
mnr-sts-gbert	0.866	7	0.864	13	0.803	4	0.998	2
mnr-sts-tsdae-gbert	0.866	7	0.929	3	0.737	9	0.998	2
mnr-jobgbert	0.854	9	0.872	12	0.790	7	0.943	9
mnr-tsdae-gbert	0.838	10	0.904	7	0.698	11	0.987	5
sts-jobgbert	0.819	11	0.877	9	0.710	10	0.919	11
tsdae-gbert	0.777	12	0.916	6	0.570	12	0.912	12
sts-tsdae-jobgbert	0.716	13	0.857	14	0.543	13	0.780	13
tsdae-jobgbert	0.676	14	0.874	10	0.516	14	0.600	14
mean	0.838		0.900		0.732		0.926	
stdev	0.069		0.032		0.113		0.110	

Table 4: Mean Average Precision (mAP@10) of the models on the random sample, evaluated on class (upper part) and concept level (lower part). Model names end with general (gbert) or domain-specific (jobgbert) LM used as starting point, each subsequent training step is prepended on the left (last step leftmost). The columns labeled 'R(ank)' denote the systems' ranking. The systems are ordered by the overall (ALL) classification performance.

	ALL	EDU	EXP	LNG
constant	0.856	0.904	0.801	0.869
MLM	-0.004	0.008	0.011	-0.058
TSDAE	-0.002	0.040	-0.030	-0.033
STS	0.024	-0.013	0.030	0.085
MNR	0.079	0.032	0.108	0.115
R2	0.616	0.348	0.733	0.643

Table 5: Linear model B-coefficients of SBERT fine-tuning steps on mAP@10 scores (class level)

Model	Context	ALL	EDU	EXP	LNG
mnr-sts-tsdae-jobgbert	all	0.908	0.923	0.865	0.963
	job ad	0.903	0.920	0.850	0.976
	ontology	0.890	0.937	0.805	0.963
	none	0.872	0.944	0.747	0.976
sts-gbert	all	0.867	0.878	0.795	0.990
	job ad	0.730	0.730	0.712	0.763
	ontology	0.852	0.901	0.735	0.990
	none	0.759	0.815	0.700	0.763

Table 6: mAP@10 for 2 selected models with different query contextualization (evaluated at concept level)

precision in a linear model (see Table 5). Over all terms, MNR raises the mAP@10 score by 7.9pp, and STS by 2.4pp, while the effects of MLM and TSDAE are small and negative.

Examining different skill types, we see that MNR is especially helpful for EXP (10.8pp) and LNG (11.5pp), much less for EDU (3.2pp). For EDU terms, the terminology is comprehensive thanks to Swiss data on educational terms, and these terms also have little class ambiguity (see Section 3.2). Thus, the smaller effect of MNR in classifying EDU terms can be explained by the fact that less needs to be learned about the ontology or the term variations. STS's strong effect on LNG (8.5pp) may reflect that this task is closer to general knowledge (e.g., *mother tongue* is similar to language proficiency), whereas EDU and EXP mapping requires domain knowledge, and barely profits from general-domain training material. TSDAE is only effective for EDU classification. Educational degrees represented in the ontology are often mentioned verbatim in job ads. We assume it is the small gap between ontology and job ad language which makes this simple fine-tuning so helpful. MLM effects are minor, but EXP classification, the most difficult task, benefits (1.1pp) from pre-training on job ad texts. In sum, MNR is the most beneficial method, but for certain term types, performance gains are observed with all approaches.

Effect of contextualization: We assess the benefits of query contextualization in ablation experiments where we omit the job ad skill span context, the ontology context, or both.[14] We compare our best model on concept level, mnr-sts-tsdae-jobgbert with sts-gbert, which has only undergone general-domain fine-tuning. Table 6 shows performance drops for both, but sts-gbert is much more affected than mnr-sts-tsdae-jobgbert (-10.8 vs -3.6pp when omitting all context). The example in Table 7 shows how mnr-sts-tsdae-jobgbert suggests appropriate skill concepts independent of contextualization, whereas sts-gbert fails without context. Examination of different term types shows that mnr-sts-tsdae-jobgbert benefits from query contextualization only for EXP mapping – the most diffi-

[14] For this ablation experiment, additional 89 skill concept suggestions were evaluated by one annotator.

Model	Context	Skill Concept	Similarity
mnr-sts-tsdae-jobgbert	all	Banking and Finance	0.722
	none	Bankier (*banker*)	0.732
sts-gbert	all	Banking Consultant	0.809
	none	Bankknecht (*butcher's assistant*)	0.782

Table 7: Most similar skill concept suggestion for the job ad expression *Bank* with and without its context *Financial Consulting, Management*

cult task –, whereas for LNG and EDU, the model seems to have incorporated enough domain knowledge during fine-tuning. As for which context is more helpful, omitting ontology context is much more detrimental to sts-gbert (-13.7pp) than omitting job ad context (-1.5pp), whereas for mnr-sts-tsdae-jobgbert, dropping job ad context is worse (-1.8 vs -0.5pp). Again, this indicates that suitable fine-tuning can effectively incorporate ontology knowledge into the model.

Low-similarity cases: We examine the 5% of EDU and EXP skills that each have the lowest similarities to ESCO concepts using mnr-sts-tsdae-jobgbert.[15] For EDU, these cases consist mainly of terms that are not skill areas at all, but containers e.g., 'Diplomabschluss' (*diploma*), rare abbreviations (*CFA (Chartered Financial Analyst)*), and generic terms like 'technisch' (*technical*).[16] For EXP, we also find mainly generic terms (*implementation*) as well as skills not covered by the ontology (e.g., *knowing a place* or *working abroad*). In a random sample of 20 low-similarity cases each for EDU and EXP, we find that for both types, 4 out of 20 skill span extractions were flawed. In the remaining cases, the precision of the first suggestion at the class level is very low, 0.594 for EDU and 0.313 for EXP. Finally, inspecting the 20 cases with the lowest similarities, none of the EDU terms and only 7 out of 20 EXP terms qualify as proper skill areas. It is in favor of our model that we find low similarities between ESCO concepts and flawed extractions or job ad skills not represented in the ontology. For practical application, the results suggest applying a minimum similarity threshold, and we use 0.5 as the default threshold.

Term selection: mAP@10 as a measure considers that the ontology may not comprise 10 acceptable suggestions for every skill area. However, for the application, a suitable cut-off value must be found for each case, since the number of acceptable ontology terms indeed varies greatly.[17] In a gradient-based approach, we aim to select term suggestions until a drop in similarities is observed, i.e., we cut off where the gradient of the probability distribution is minimal. This way, we consider on average three ontology terms for each skill term. In comparison to considering only the most similar term, we lose on average 3.6pp of the evaluation score on the concept level, which we regard as a reasonable trade-off for application.

6 Downstream Sociological Analysis

Labor market changes: It is commonly believed that digital technologies have changed the demand for skills to perform tasks in the labor market in the past decades. Recent literature points to the importance of new skills entering jobs and altering the required skill combinations (Acemoglu et al., 2022). It also emphasizes that most of the changes in skill demand take place within and not across occupations (Bisello et al., 2019; Freeman et al., 2020). According analyses require time series data on skill demand at the job level that includes valid measures of all skills required in the labor market. Such data has been, however, extremely scarce.

Illustrative analyses: To illustrate the usefulness of our job ads data for social sciences, we present some selected analyses. First, we calculate correlations between occupation-skill matrices that ESCO provides and those resulting from the SJMM data. At the 1-digit level of the international standard classification of occupations (ISCO-08)[18], for example, the correlation is as high as 0.87, underscoring the validity of our skill extractions. Second, we illustrate within-occupation change in skill demand with an example: the evolution of skill requirements in the occupational field of technicians and engineers. To aggregate fine-grained, multi-hierarchical ESCO skill classes, we used a clustering approach.[19] The resulting 48 clusters are then applied to the SJMM job level data, generating for each job ad indicators of how strongly the text represents each skill cluster. To keep the picture detailed as well as simple, only three interesting clusters for this occupation are shown in Figure 3.

Figure 3 confirms – for our example – that

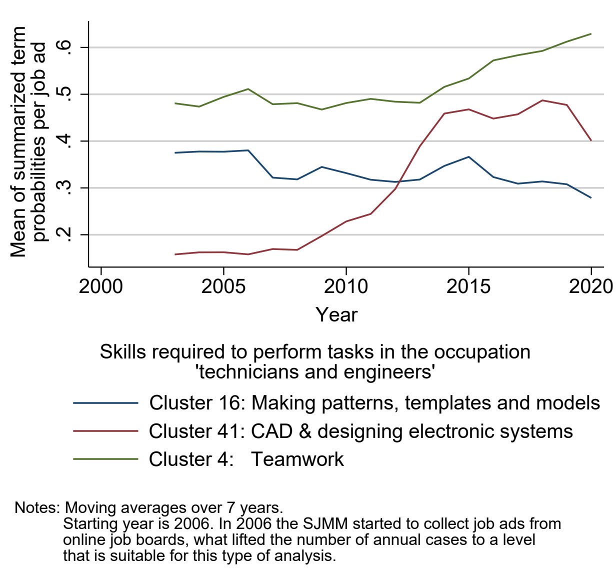
[15] Mean similarity of the first suggestion is 0.446 for EXP and 0.473 for EDU.

[16] The German word 'technisch' (*technical*) appears in 377 ontology terms or descriptions, from 183 different classes.

[17] For instance, five concepts were accepted for *acquiring new customers*, but only one for *fire department*.

[18] https://isco-ilo.netlify.app/en/isco-08/

[19] We applied HDBSCAN (Campello et al., 2013) (min. size=3, epsilon=0.0 and alpha=1.0) over skill class vectors (averaged skill term vectors per class).



Notes: Moving averages over 7 years.
Starting year is 2006. In 2006 the SJMM started to collect job ads from online job boards, what lifted the number of annual cases to a level that is suitable for this type of analysis.

Figure 3: Illustration of within-occupation evolution of skill requirements

the type of required skills changed within occupations over time. Skills for *making patterns, templates and models* were highly required shortly after the turn of the century. Across the following 15 years, the demand for these mainly manual and non-digital skills declined. In contrast, the demand for *CAD and designing electronic systems* was nearly nonexistent and then increased sharply. These skills are related to digital technologies and newly entered the occupation. After their entry, also other elements of the required skill combination seem to change, e.g., demand for *teamwork skills* is increasing (see Figure 3). This is in line with the literature, which suggests that digital technologies lead to more flexible, team-based settings (Autor et al., 2002).

7 Conclusion

Our two-step approach of first extracting text spans expressing language skills, experience, and educational requirements, followed by further subdividing these into skill areas, containers, and qualifiers, allowed us to achieve broad coverage of fine-grained competency classifications. By grouping skill areas from the same span for transformer-based vector representation, we provide relevant context that helps find appropriate ESCO ontology concepts for each job ad skill area.

For fine-grained classification, our domain and task-specific SBERT learning steps boost performance – best models reaching mAP@10 of 0.969 on class and 0.908 on concept level – and also help deal with more difficult cases encountered in the challenge sample. While infusing terminological variation from the ontology into the model with MNR is by far the most effective, all different pretraining and fine-tuning steps are beneficial to some extent.

Analyses on low-similarity cases and our gradient-based selection approach showed that similarity values of our best models can be used to select the most relevant ontology concepts and avoid mismatches.

In future work, models could be further fine-tuned with curated task-specific training material (similar to our evaluation data) to improve classification for the most difficult task, experience classification (EXP). The next steps in social science analyses could be to assess how required skill combinations evolve within occupations, which occupations shift towards more specialized or diversified requirements, or to which extent the skill requirements of some occupations become more alike.

Limitations

Job ad texts are influenced by conventions, social norms, and the effects of their publication media. This potentially affects the performance of our approach in different social settings, e.g., for German-language job ads from other countries.

Furthermore, the average number of skill requirements per ad grows over time. The extent to which this is due to changes in labor market structure, social norms, recruiting practices, or publication media remains to be investigated.

Our SBERT fine-tuning aimed at enabling valid skill classification for job ads from the last three decades. Therefore, the application to future job ads might require periodic updates of models with newer data. And, while our experiments on the classification task show expected and explainable results, analyses could still benefit from a larger test set.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by the Swiss National Science Foundation (grant number 407740 187333).

References

Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. Artificial intelligence and jobs: evidence from online vacancies. *Journal of Labor Economics*, 40(S1):S293–S340.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2020. The Evolution of Work in the United States. *American Economic Journal: Applied Economics*, 12(2):1–34.

David H. Autor, Frank Levy, and Richard J. Murnane. 2002. Upstairs, downstairs: Computers and skills on two floors of a large bank. *ILR Review*, 55(3):432–447.

José A. Azar, Ioana Marinescu, Marshall I. Steinbaum, and Bledi Taska. 2018. Concentration in US Labor Markets: Evidence From Online Vacancy Data.

Federico Biagi and Raquel Sebastian. 2020. Technologies and "routinization". *Handbook of Labor, Human Resources and Population Economics*, pages 1–17.

Martina Bisello, Eleonora Peruffo, Enrique Fernández-Macías, and Riccardo Rinaldi. 2019. How computerisation is transforming jobs: Evidence from the eurofound's european working conditions survey. Technical report, JRC working papers series on Labour, Education and Technology.

Marlis Buchmann, Helen Buchs, Felix Busch, Simon Clematide, Ann-Sophie Gnehm, and Jan Müller. 2022a. Swiss Job Market Monitor: A Rich Source of Demand-Side Micro Data of the Labour Market. *European Sociological Review*.

Marlis Buchmann, Helen Buchs, Eva Bühlmann, Felix Busch, Ann-Sophie Gnehm, Yanik Kipfer, Urs Klarer, Jan Müller, Marianne Müller, Stefan Sacchi, Alexander Salvisbert, and Anna von Ow. 2022b. *Stellenmarkt-Monitor Schweiz 1950 – 2021.* Soziologisches Institut der Universität Zürich.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Deming and Lisa B Kahn. 2018. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Commission. Directorate General for Employment, Social Affairs and Inclusion. 2017. *ESCO handbook: European skills, competences, qualifications and occupations.* Publications Office, LU.

Richard B. Freeman, Ina Ganguli, and Michael J. Handel. 2020. Within-occupation changes dominate changes in what workers do: A shift-share decomposition, 2005–2015. *AEA Papers and Proceedings*, 110:394–99.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Ann-Sophie Gnehm, Eva Bühlmann, and Simon Clematide. 2022. Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3892–3901, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. ArXiv:1705.00652 [cs].

Brad Hershbein and Lisa B Kahn. 2018. Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, 108(7):1737–72.

Klaus Krippendorff. 2004. Reliability in Content Analysis.: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press, New York. OCLC: ocn190786122.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. 2022a. SkillSpan: Hard and Soft Skill Extraction from English Job Postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4962–4984, Seattle, United States. Association for Computational Linguistics.

Mike Zhang, Kristian Nørgaard Jensen, and Barbara Plank. 2022b. Kompetencer: Fine-grained skill classification in danish job postings via distant supervision and transfer learning. In *Proceedings of the Language Resources and Evaluation Conference*, pages 436–447, Marseille, France. European Language Resources Association.