# ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain

**Mike Zhang** and **Rob van der Goot** and **Barbara Plank**
Department of Computer Science, IT University of Copenhagen, Denmark
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
{mikz, robv}@itu.dk b.plank@lmu.de

## Abstract

The increasing number of benchmarks for Natural Language Processing (NLP) tasks in the computational job market domain highlights the demand for methods that can handle job-related tasks such as skill extraction, skill classification, job title classification, and de-identification. While some approaches have been developed that are specific to the job market domain, there is a lack of generalized, multilingual models and benchmarks for these tasks. In this study, we introduce a language model called ESCOXLM-R, based on XLM-$R_{large}$, which uses domain-adaptive pre-training on the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy, covering 27 languages. The pre-training objectives for ESCOXLM-R include dynamic masked language modeling and a novel additional objective for inducing multilingual taxonomical ESCO relations. We comprehensively evaluate the performance of ESCOXLM-R on 6 sequence labeling and 3 classification tasks in 4 languages and find that it achieves state-of-the-art results on 6 out of 9 datasets. Our analysis reveals that ESCOXLM-R performs better on short spans and outperforms XLM-$R_{large}$ on entity-level and surface-level span-F1, likely due to ESCO containing short skill and occupation titles, and encoding information on the entity-level.

## 1 Introduction

The dynamic nature of labor markets, driven by technological changes, migration, and digitization, has resulted in a significant amount of job advertisement data (JAD) being made available on various platforms to attract qualified candidates (Brynjolfsson and McAfee, 2011, 2014; Balog et al., 2012). This has led to an increase in tasks related to JAD, including skill extraction (Kivimäki et al., 2013; Zhao et al., 2015; Sayfullina et al., 2018; Smith et al., 2019; Tamburri et al., 2020; Shi et al., 2020; Chernova, 2020; Bhola et al., 2020; Zhang et al., 2022a,b,c; Green et al., 2022; Gnehm et al., 2022; Beauchemin et al., 2022; Decorte et al., 2022; Goyal et al., 2023), skill classification (Decorte et al., 2022; Zhang et al., 2022b), job title classification (Javed et al., 2015, 2016; Decorte et al., 2021; Green et al., 2022), de-identification of entities in job postings (Jensen et al., 2021), and multilingual skill entity linking (ESCO, 2022).

While some previous studies have focused on JAD in non-English languages (Zhang et al., 2022b; Gnehm et al., 2022; Beauchemin et al., 2022), their baselines have typically relied on language-specific models, either using domain-adaptive pre-training (DAPT; Gururangan et al., 2020) or off-the-shelf models. The lack of comprehensive, open-source JAD data in various languages makes it difficult to fully pre-train a language model (LM) using such data. In this work, we seek external resources that can help improve the multilingual performance on the JAD domain. We use the ESCO taxonomy (le Vrang et al., 2014), which is a standardized system for describing and categorizing the skills, competences, qualifications, and occupations of workers in the European Union. The ESCO taxonomy, which has been curated by humans, covers over 13,000 skills and 3,000 occupations in 27 languages. Therefore, we seek to answer: *To what extent can we leverage the ESCO taxonomy to pre-train a domain-specific and language-agnostic model for the computational job market domain?*

In this work, we release the first multilingual JAD-related model named ESCOXLM-R, a language model based on XLM-$R_{large}$ that incorporates data from the ESCO taxonomy through the use of two pre-training objectives (Figure 1): Masked Language Modeling (MLM) and a novel ESCO relation prediction task (Section 2). We evaluate ESCOXLM-R on 9 JAD-related datasets in 4 different languages covering 2 NLP tasks (Section 3). Our results show that ESCOXLM-R outperforms previous state-of-the-art (SOTA) on 6 out of 9 datasets (Section 4). In addition, our fine-grained



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 11871–11890

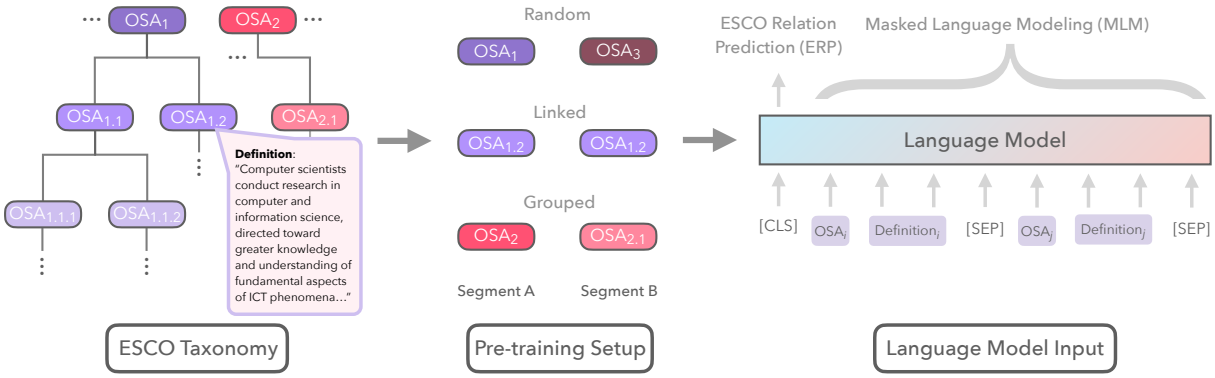


Figure 1: **ESCO Pre-training Objective**: From left to right, the figure illustrates the hierarchical structure of the ESCO taxonomy, which consists of occupations, skills, and aliases (OSA). Each OSA includes a definition. For the purposes of this study, we consider aliases of occupations to have the same definition as the occupation itself. In the middle of the figure, we show our pre-training setup. Pre-training instances are uniformly sampled in three ways: randomly, linked, or grouped (this is defined in Section 2.2). The selected instances (can be in different languages) are then fed to the language model, along with its description. We have two pre-training objectives: the regular MLM objective, and a new ESCO relation prediction objective, in which the goal is to predict which group the sampled instances belong to (Random, Linked, or Grouped).

analysis reveals that ESCOXLM-R performs better on short spans compared to XLM-R$_{\text{large}}$, and consistently outperforms XLM-R$_{\text{large}}$ on entity-level and surface-level span-F1 (Section 5).

**Contributions** In this work, we present and release the following:

- ESCOXLM-R, an XLM-R$_{\text{large}}$-based model, which utilizes domain-adaptive pre-training on the 27 languages from ESCO.[1]
- The largest JAD evaluation study to date on 3 job-related tasks, comprising 9 datasets in 4 languages and 4 models.
- A fine-grained analysis of ESCOXLM-R's performance on different span lengths, and emerging entities (i.e., recognition of entities in the long tail).

## 2 ESCOXLM-R

**Preliminaries** In the context of pre-training, an LM is trained using a large number of unlabeled documents, $\mathcal{X} = X^{(i)}$, and consists of two main functions: $f_{\text{encoder}}(.)$, which maps a sequence of tokens $X = (x_1, x_2, ..., x_t)$ to a contextualized vector representation for each token, represented as $(h_1, h_2, ..., h_t)$, and $f_{\text{head}}(.)$, the output layer that takes these representations and performs a specific task, such as pre-training in a self-supervised manner or fine-tuning on a downstream application. For example, BERT (Devlin et al., 2019) is pre-trained using two objectives: MLM and Next Sentence Prediction (NSP). In MLM, a portion of tokens in a sequence $X$ is masked and the model must predict the original tokens from the masked input. In the NSP objective, the model takes in two segments $(X_{\text{A}}, X_{\text{B}})$ and predicts whether segment $X_{\text{B}}$ follows $X_{\text{A}}$. RoBERTa (Liu et al., 2019) is a variation of BERT that uses dynamic MLM, in which the masking pattern is generated each time a sequence is fed to the LM, and does not use the NSP task.

**Multilinguality** Both BERT and RoBERTa have been extended to support multiple languages, resulting in multilingual BERT (mBERT; Devlin et al., 2019) and XLM-RoBERTa (XLM-R; Conneau et al., 2020). XLM-R was found to outperform mBERT on many tasks (e.g., Conneau et al., 2020; Hu et al., 2020; Lauscher et al., 2020) due to careful tuning, sampling, and scaling to larger amounts of textual data. Because of this, our ESCOXLM-R model is based on XLM-R$_{\text{large}}$.

### 2.1 European Skills, Competences, Qualifications and Occupations Taxonomy

The European Skills, Competences, Qualifications, and Occupations (ESCO; le Vrang et al., 2014) taxonomy is a standardized system for describing and categorizing the skills, competences, qualifications, and occupations of workers in the European Union (EU). It is designed to serve as a common lan-

[1]The code for ESCOXLM-R is available as open-source: https://github.com/mainlp/escoxlmr. We further release ESCOXLM-R under an Apache License 2.0 on HuggingFace: https://huggingface.co/jjzha/esco-xlm-roberta-large.

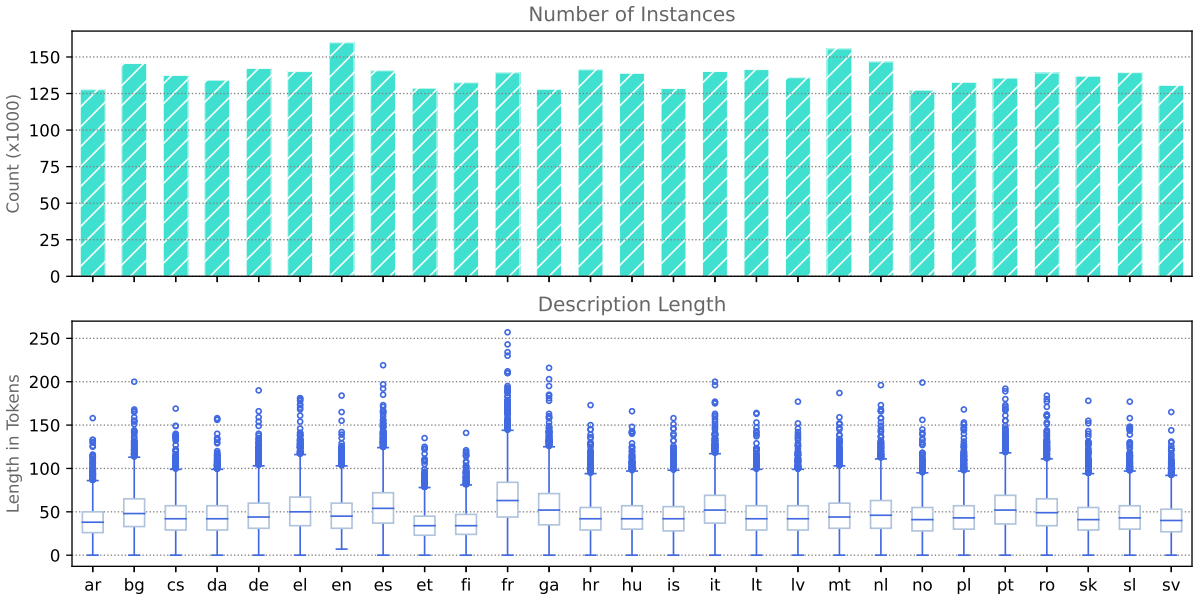


Figure 2: **Statistics of Pre-training Data.** The ESCO dataset contains descriptions in 27 languages, with a combined total of approximately 3.72 million descriptions (i.e., instances). On average, there are around 130,000 descriptions per language. The average length of each description is 26.3 tokens, with some descriptions reaching a maximum length of 150 or more tokens, as shown by the outliers in the boxplot.

guage for the description of skills and qualifications across the EU, facilitating the mobility of workers by providing a common reference point for the recognition of qualifications and occupations. The taxonomy is developed and maintained by the European Commission and is based on the International Classification of Occupations and the International Standard Classification of Education. It includes 27 European languages: Bulgarian (ar), Czech (cs), Danish (da), German (de), Greek (el), English (en), Spanish (es), Estonian (et), Finnish (fi), French (fr), Gaelic (ga), Croatian (hr), Hungarian (hu), Icelandic (is), Italian (it), Lithuanian (lt), Latvian (lv), Maltese (mt), Dutch (nl), Norwegian (no), Polish (pl), Portuguese (pt), Romanian (ro), Slovak (sk), Slovenian (sl), Swedish (sv), and Arabic (ar). Currently, it describes 3,008 occupations and 13,890 skills/competences (SKC) in all 27 languages.[2]

The ESCO taxonomy includes a hierarchical structure with links between occupations, skills, and aliases (OSA). In this work, we focus on the occupation pages and extract the following information from the taxonomy:[3]

- `ESCO Code`: The taxonomy code for the specific occupation or SKC.
- `Occupation Label`: The preferred occupation name (i.e., title of the occupation).
- `Occupation Description/Definition`: A description of the responsibilities of the specific occupation.
- `Major Group Name`: The name of the overarching group to which the occupation belongs, e.g., "Veterinarians" for the occupation "animal therapist".
- `Alternative Labels`: Aliases for the specific occupation, e.g., "animal rehab therapist" for the occupation "animal therapist".
- `Essential Skills`: All necessary SKCs for the occupation, including descriptions of these.
- `Optional Skills`: All optional SKCs for the occupation, including descriptions of these.

In Figure 2, we present the distribution of pre-training instances and the mean description lengths for each language in the ESCO taxonomy. Note that the number of descriptions is not the same for all languages, and we do not count empty descriptions (i.e., missing translations) for certain occupations or SKCs.

[2]Note that ESCO now also includes Ukrainian, but this model was trained before that inclusion. We use the ESCO V1.0.9 API to extract the data. ESCO contains an Apache 2.0 and a European Union Public License 1.2.

[3]An example of the extracted information can be found in Listing 1 (Appendix A), and the original page can be accessed at `https://bit.ly/3DY1zsX`.

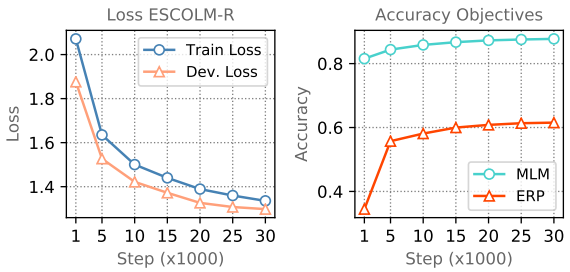


Figure 3: **Pre-Training Statistics.** The final log loss for the training set is 1.34, while the log loss for the development set is 1.30. The MLM accuracy is 84.3%, while the Entity Relationship Prediction (ERP) accuracy is 60.0%. These results were obtained after approximately 1.04 epochs of training on the total data.

## 2.2 Pre-training Setup

To improve our XLM-R$_{\text{large}}$-based model, we employ domain-adaptive pre-training techniques as described in previous work such as Alsentzer et al. (2019); Han and Eisenstein (2019); Lee et al. (2020); Gururangan et al. (2020); Nguyen et al. (2020). Given the limited amount of training data (3.72M sentences), we utilize the XLM-R$_{\text{large}}$ checkpoint provided by the HuggingFace library (Wolf et al., 2020) as a starting point.[4] Our aim is to fine-tune the model to internalize domain-specific knowledge related to occupation and SKCs, while maintaining its general knowledge acquired during the original pre-training phase.

We introduce a novel self-supervised pre-training objective for ESCOXLM-R, inspired by LinkBERT from Yasunaga et al. (2022). We view the ESCO taxonomy as a graph of occupations and SKCs (Figure 1), with links between occupations or occupations and SKCs in various languages. By placing similar occupations or SKCs in the same context window and in different languages, we can learn from the links between (occupation $\leftrightarrow$ occupation) and (occupation $\leftrightarrow$ SKCs) in different languages for true cross-lingual pre-training. In addition to the MLM pre-training objective, which is used to learn concepts within contexts, we introduce another objective called ESCO Relation Prediction (ERP) to internalize knowledge of connections within the taxonomy in the LM. We take an anchor concept ($C_{\text{A}}$) by concatenating it with its description ($X_{\text{A}}$) from the ESCO taxonomy and sample an additional concept ($C_{\text{B}}$) concatenated with its description ($X_{\text{B}}$) to create LM input [CLS] $C_{\text{A}}X_{\text{A}}$ [SEP] $C_{\text{B}}X_{\text{B}}$ [SEP].[5] We sample $C_{\text{B}}X_{\text{B}}$ in three ways with uniform probability:

1. *Random*: We randomly sample $C_{\text{B}}X_{\text{B}}$ from the ESCO taxonomy, in any language;
2. *Linked*: We sample $C_{\text{B}}X_{\text{B}}$ in any language from the same occupation page, for example, an "animal therapist" (or an alias of the "animal therapist", e.g., "animal rehab therapist") should have knowledge of "animal behavior";
3. *Grouped*: We sample $C_{\text{B}}X_{\text{B}}$ from the same major group in any language. For the same example "animal therapist", it comes from major group 2: Professionals $\rightarrow$ group 22: Health professionals. Several other concepts, e.g., "Nursing professionals" fall under this major group.

**Pre-training Objectives** The LM is trained using two objectives. First is the MLM objective, and the second is the ERP objective, where the task is to classify the relation $r$ of the [CLS] token in [CLS] $C_{\text{A}}X_{\text{A}}$ [SEP] $C_{\text{B}}X_{\text{B}}$ [SEP] ($r \in$ Random, Linked, Grouped). The rationale behind this is to encourage the model to learn the relevance between concepts in the ESCO taxonomy. We formalize the objectives in Equation (1):

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ERP}} \\ &= -\sum_i \log p\left(x_i \mid \mathbf{h}_i\right) - \log p\left(r \mid \mathbf{h}_{\texttt{[CLS]}}\right),\end{aligned} \tag{1}$$

we define the overall loss $\mathcal{L}$ as the sum of the MLM loss $\mathcal{L}_{\text{MLM}}$ and the ERP loss $\mathcal{L}_{\text{ERP}}$. The MLM loss is calculated as the negative log probability of the input token $x_i$ given the representation $\mathbf{h}_i$. Similarly, the ERP loss is the negative log probability of the relationship $r$ given the representation of the start-token $\mathbf{h}_{\texttt{[CLS]}}$. In our implementation, we use XLM-R$_{\text{large}}$ and classify the start-token [CLS] for ERP to improve the model's ability to capture the relationships between ESCO occupations and skills.

[4] https://huggingface.co/xlm-roberta-large

[5] The special tokens used in this example follow the naming convention of BERT for readability, [CLS] and [SEP]. However, since we use XLM-R$_{\text{large}}$ there are different special tokens: <s> as the beginning of the sequence, </s> as the SEP token, and </s></s> as segment separators. Formally, given the example in the text: <s> $C_{\text{A}}X_{\text{A}}$ </s></s> $C_{\text{B}}X_{\text{B}}$ </s>.

| Dataset Name | Lang. | Loc. | License | Task | Metric | Input Type | Train | Dev. | Test |
|---|---|---|---|---|---|---|---|---|---|
| SKILLSPAN | en | * | CC-BY-4.0 | SL | Span-F1 | Sentences | 5,866 | 3,992 | 4,680 |
| SAYFULLINA | en | UK | Unknown | SL | Span-F1 | Sentences | 3,706 | 1,854 | 1,853 |
| GREEN | en | UK | CC-BY-4.0 | SL | Span-F1 | Sentences | 8,670 | 963 | 336 |
| JOBSTACK | en | * | RLT | SL | Span-F1 | Sentences | 18,055 | 2,082 | 2,092 |
| BHOLA | en | SG | CC-BY-4.0 | MLC | MRR | Documents | 16,238 | 2,030 | 2,030 |
| KOMPETENCER | en | DK | CC-BY-4.0 | MCC | W. Macro-F1 | Skills | 9,472 | 1,577 | 1,578 |
| KOMPETENCER | **da** | DK | CC-BY-4.0 | MCC | W. Macro-F1 | Skills | 138 | - | 784 |
| GNEHM | **de** | CH | CC-BY-NC-SA-4.0 | SL | Span-F1 | Sentences | 22,134 | 2,679 | 2,943 |
| FIJO | **fr** | FR | Unknown | SL | Span-F1 | Sentences | 399 | 50 | 50 |

Table 1: **Dataset Statistics.** We show statistics for all 9 JAD datasets. There are 6 datasets in English and 3 in other languages (Danish, German, and French). We indicate the location the JAD originates from (whenever applicable, * indicates it comes from a variety of countries). We indicate the license of the dataset. Most of the task types consist of sequence labeling (e.g., span extraction, Named Entity Recognition, soft skill tagging). To maintain consistency, we use a single metric for each task type: Sequence Labeling (SL), Multilabel Classification (MLC), and Multiclass Classification (MCC). For KOMPETENCER, the statistics are provided in brackets for the Danish language.

**Implementation** For optimization we follow (Yasunaga et al., 2022), we use the AdamW (Loshchilov and Hutter, 2019) optimizer with ($\beta_1$, $\beta_2$) = (0.9, 0.98). We warm up the learning rate $1e^{-5}$ for a ratio of 6% and then linearly decay it. The model is trained for 30K steps, which is equivalent to one epoch over the data, and the training process takes 33 hours on one A100 GPU with tf32. We use a development set comprising 1% of the data for evaluation. In Figure 3, the pre-training loss and performance on the dev. set are plotted, it can be seen that the accuracy plateaus at 30K steps. Though the train and development loss hint that further gains could be obtained on the pretraining objective, we found through empirical analysis on downstream tasks that 30K steps performs best.

## 3 Experimental Setup

Table 1 provides the details of the downstream datasets used in this study. Most of the datasets are in EN, with a smaller number in DA, DE, and FR. For each dataset, a brief description and the corresponding best-performing models are given. We put examples of each dataset (apart from JobStack due to the license) in Appendix B.

**SKILLSPAN (Zhang et al., 2022a)** The job posting dataset includes annotations for skills and knowledge, derived from the ESCO taxonomy. The best model in the relevant paper, JobBERT, was retrained using a DAPT approach on a dataset of 3.2 million EN job posting sentences. This is the best-performing model which we will compare against.

**KOMPETENCER (Zhang et al., 2022b)** This dataset is used to evaluate models on the task of classifying skills according to their ESCO taxonomy code. It includes EN and DA splits, with the EN set derived from SKILLSPAN. There are three experimental setups for evaluation: fully supervised with EN data, zero-shot classification (EN→DA), and few-shot classification (a few DA instances). The best-performing model in this work is RemBERT (Chung et al., 2021), which obtains the highest weighted macro-F1 for both EN and DA. In this work, we use setup 1 and 3, where all available data is used.

**BHOLA (Bhola et al., 2020)** The task of this EN job posting dataset is multilabel classification: Predicting a list of necessary skills in for a given job description. It was collected from a Singaporean government website. It includes job requirements and responsibilities as data fields. Pre-processing steps included lowercasing, stopword removal, and rare word removal. Their model is BERT with a bottleneck layer (Liu et al., 2017). In our work, the bottleneck layer is not used and no additional training data is generated through bootstrapping. To keep comparison fair, we re-train their model without the additional layer and bootstrapping. We use Mean Reciprocal Rank as the main results metric.

**SAYFULLINA (Sayfullina et al., 2018)** This dataset is used for soft skill prediction, a sequence labeling problem. Soft skills are personal qualities that contribute to success, such as "team working", "being dynamic", and "independent". The models for this dataset include a CNN (Kim, 2014), an LSTM (Hochreiter et al., 1997), and a Hierarchical Attention Network (Yang et al., 2016). We compare to their best-performing LSTM model.

| Dataset | Lang. | Metric | Prev. SOTA | XLM-R$_{large}$ | XLM-R$_{large}$ (+ DAPT) | ESCOXLM-R | Δ |
|---|---|---|---|---|---|---|---|
| SKILLSPAN | EN | Span-F1 | 58.9±4.5 | 59.7±4.6 | 62.0±4.0 | **62.6±3.7** | +3.7 |
| SAYFULLINA | EN | Span-F1 | 73.1±2.1 | 89.9±0.5 | 90.6±0.4 | **92.2±0.2** | +19.1 |
| GREEN | EN | Span-F1 | 31.8±* | 49.0±2.4 | 47.5±0.7 | **51.2±2.1** | +19.4 |
| JOBSTACK | EN | Span-F1 | **82.1±0.8** | 81.2±0.6 | 80.4±0.7 | 82.0±0.7 | −0.1 |
| KOMPETENCER | EN | W. Macro-F1 | 62.8±2.8 | 59.0±9.5 | **64.3±0.5** | 63.5±1.3 | −0.7 |
| BHOLA | EN | MRR | 90.2±0.2 | 90.5±0.3 | 90.0±0.3 | **90.7±0.2** | +0.5 |
| GNEHM | DE | Span-F1 | 86.7±0.4 | 87.1±0.4 | 86.8±0.2 | **88.4±0.5** | +1.7 |
| FIJO | FR | Span-F1 | 31.7±2.3 | 41.8±2.0 | 41.7±0.7 | **42.0±2.3** | +10.3 |
| KOMPETENCER | DA | W. Macro-F1 | 45.3±1.5 | 41.2±9.8 | **45.6±0.8** | 45.0±1.4 | −0.3 |

Table 2: **Results of Experiments.** The datasets and models are described in Section 3. We re-train the best-performing models of all papers to give us the standard deviation. The best-performing model is in bold. The difference in performance between ESCOXLM-R and the previous SOTA is shown as Δ. Note (*) that the results for GREEN are based on a CRF model where the data has been pre-split, and therefore, there is no standard deviation.

**GREEN (Green et al., 2022)** A sentence-level sequence labeling task involving labeling skills, qualifications, job domain, experience, and occupation labels. The job positions in the dataset are from the United Kingdom. The industries represented in the data vary and include IT, finance, healthcare, and sales. Their model for this task is a Conditional Random Field (Lafferty et al., 2001) model.

**JOBSTACK (Jensen et al., 2021)** This corpus is used for de-identifying personal data in job vacancies on Stack Overflow. The task involves sequence labeling and predicting Organization, Location, Name, Profession, and Contact details labels. The best-performing model for this task is a transformer-based (Vaswani et al., 2017) model trained in a multi-task learning setting. Jensen et al. (2021) propose to use the I2B2/UTHealth corpus, which is a medical de-identification task (Stubbs and Uzuner, 2015), as auxiliary data, which showed improvement over their baselines.

**GNEHM (Gnehm et al., 2022)** A Swiss-German job ad dataset where the task is Information and Communications Technology (ICT)-related entity recognition, these could be ICT tasks, technology stack, responsibilities, and so forth. The used dataset is a combination of two other Swiss datasets namely the Swiss Job Market Monitor and an online job ad dataset (Gnehm and Clematide, 2020; Buchmann et al., 2022). Their model is dubbed JobGBERT and is based on DAPT with German BERT$_{base}$ (Chan et al., 2020).

**FIJO (Beauchemin et al., 2022)** A French job ad dataset with the task of labeling skill types using a sequence labeling approach. The skill groups are based on the AQESSS public skills repositories and proprietary skill sets provided by their collaborators. These skill types are divided into four categories: "Thoughts", "Results", "Relational", and "Personal". The best-performing model for this task is CamemBERT (Martin et al., 2020).

## 4 Results

The results of the models are presented in Table 2. To evaluate the performance, four different models are used in total: ESCOXLM-R, the best-performing model originally reported in the relevant paper for the downstream task, vanilla XLM-R$_{large}$, and an XLM-R$_{large}$ model that we continuously pre-trained using only MLM (DAPT; excluding the ERP objective) using the same pre-training hyperparameters as ESCOXLM-R. For more information regarding the hyperparameters of fine-tuning, we refer to Appendix C (Table 5).

**English** ESCOXLM-R is the best-performing model in 4 out of 6 EN datasets. The largest improvement compared to the previous SOTA is observed in SAYFULLINA and GREEN, with over 19 F1 points. In 3 out of 4 datasets, ESCOXLM-R has the overall lower standard deviation. For JOBSTACK, the previous SOTA performs best, and for KOMPETENCER, XLM-R$_{large}$ (+ DAPT) has the highest performance.

**Non-English** In 2 out of 3 datasets, ESCOXLM-R improves over the previous SOTA, with the largest absolute difference on French FIJO with 10.3 F1 points. In the Danish subset of KOMPETENCER, XLM-R$_{large}$ (+ DAPT) has higher performance than ESCOXLM-R. Next, we will discuss potential reasons for these differences.

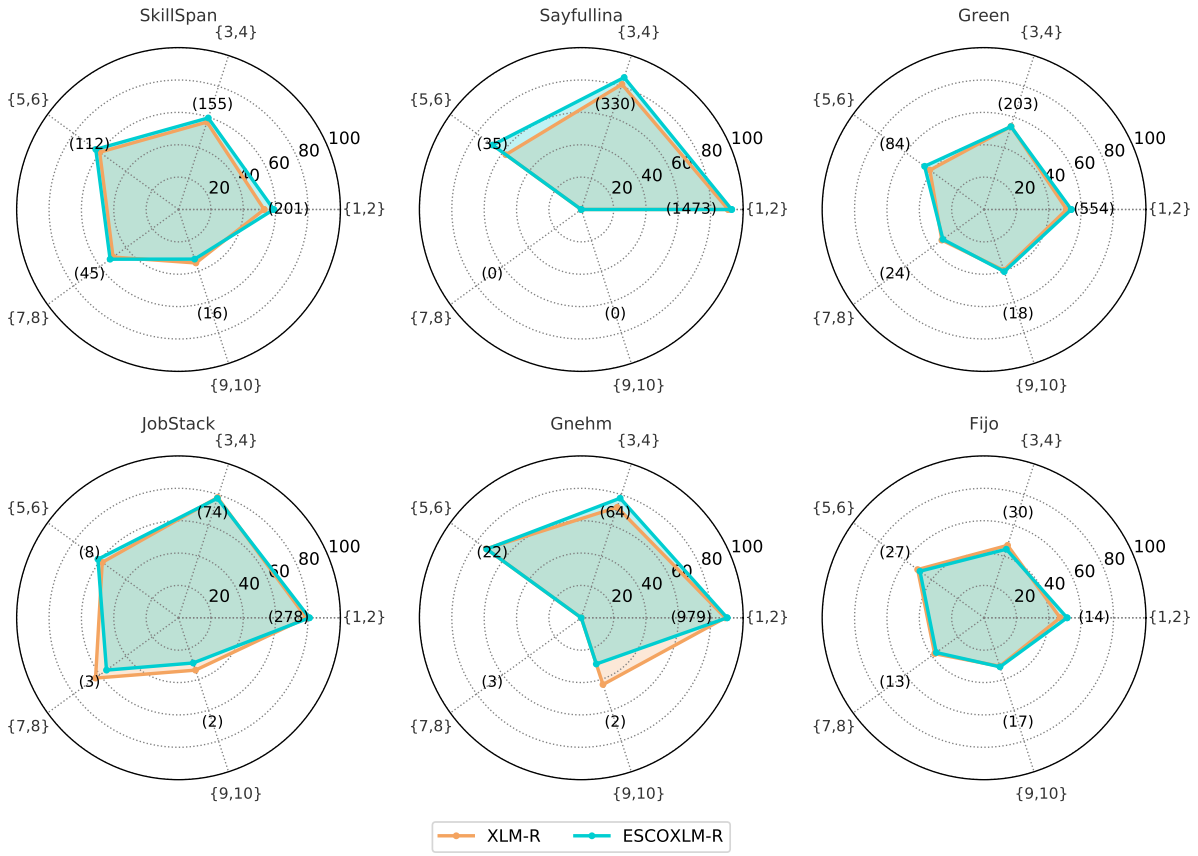


Figure 4: **Radar Charts of Span-F1 performance by Span Token Length.** We show the performance of XLM-$R_{large}$ and ESCOXLM-R on different span lengths, we bucketed the performances of both models according to the length of the spans, up to 10 tokens, and presented the average performance over five random seeds. We did not include error bars in these plots. Note that in some plots, there are no instances in certain buckets (e.g., SAYFULLINA with 7-8, 9-10). Also, some outer rings only go up to 60 span F1, rather than 100.

### 4.1 Analysis

We highlight that the performance gains of ESCOXLM-R are generally much larger than any of the losses, indicating a largely positive effect of training on ESCO. The improved performance of ESCOXLM-R on JAD datasets in Table 2 is likely due to the focus on tasks with token-level annotation (i.e., sequence labeling). This suggests that pre-training on the ESCO taxonomy is particularly useful for these types of tasks. The under-performance of ESCOXLM-R on the KOMPETENCER dataset in both EN and DA may be because the task involves predicting the ESCO taxonomy code for a given skill *without context*, where we expect ESCO to particularly help with tasks where having context is relevant. We suspect applying DAPT and ERP on ESCO specifically improves recognizing entities that are uncommon. On the other hand, the poor performance on the JOBSTACK dataset may be due to the task of predicting various named entities, such as organizations and locations. By manual inspection, we found that ESCO does not contain entities related to organizations, locations, or persons, thus this reveals that there is a lack of relevant pre-training information to JOBSTACK.

## 5 Discussion

### 5.1 Performance on Span Length

We seek to determine whether the difference in performance between the ESCOXLM-R and XLM-$R_{large}$ models is due to shorter spans, and to what extent. One application of predicting short spans well is the rise of technologies, for which the names are usually short in length. Zhang et al. (2022c) observes that skills described in the ESCO dataset are typically short, with a median length of approximately 3 tokens. We compare the average performance of both models on the test sets of each dataset, where span-F1 is used as measurement. We group gold spans into buckets of lengths 1-2, 3-4, 5-6, 7-8,

| Dataset | Ratio | Span-F1 (Entity) | | Span-F1 (Surface) | |
|---|---|---|---|---|---|
| | | XLM-R | ESCOXLM-R | XLM-R | ESCOXLM-R |
| SKILLSPAN | 0.90 | 59.9±7.9 | **61.6±6.6** | 56.4±5.7 | **57.9±4.3** |
| SAYFULLINA | 0.22 | 94.0±0.2 | **95.7±0.3** | 82.8±0.6 | **87.2±0.7** |
| GREEN | 0.79 | 50.3±2.4 | **53.1±2.1** | 49.2±2.4 | **52.0±2.1** |
| JOBSTACK | 0.41 | 85.6±0.7 | **86.4±0.5** | 78.4±1.2 | **79.8±0.7** |
| GNEHM | 0.53 | 89.3±0.3 | **89.6±0.4** | 87.3±0.3 | **87.8±0.6** |
| FIJO | 0.77 | 34.4±2.9 | **35.7±1.1** | 34.4±1.1 | **35.7±1.1** |

Table 3: **Entity vs. Surface-level span-F1 on Test.** In this table, the performance of two systems, XLM-$R_{large}$ and ESCOXLM-R, was measured using entity-level and surface-level span-F1 scores. Entity-level span-F1 measures precision, recall, and harmonic mean at the entity level, while surface-level span-F1 measures a system's ability to recognize a range of entities. We include the ratio of surface entities to total entities in each *training* set, with a higher ratio indicating more variety (a ratio of 1.00 indicates all entities are unique).

and 9-10, and present the span-F1 for each model (XLM-$R_{large}$ vs. ESCOXLM-R) in each bucket.

Shown in Figure 4, ESCOXLM-R outperforms XLM-$R_{large}$ on shorter spans (i.e., 1-2 or 3-4) in 6 out of the 6 datasets, suggesting that pre-training on ESCO is beneficial for predicting short spans. However, there is a slight decline in performance on some datasets (e.g., SKILLSPAN, JOBSTACK, and GNEHM) when the spans are longer (i.e., 7-8 or 9-10). It is worth noting that the number of instances in these longer span buckets is lower, and therefore errors may be less apparent in terms of their impact on overall performance.

### 5.2 Entity-F1 vs. Surface-F1

In this analysis, we adopt the evaluation method used in the W-NUT shared task on Novel and Emerging Entity Recognition (Derczynski et al., 2017). In this shared task, systems are evaluated using two measures: entity span-F1 and surface span-F1. Entity span-F1 assesses the precision, recall, and harmonic mean (F1) of the systems at the entity level, while surface span-F1 assesses their ability to correctly recognize a diverse range of entities, rather than just the most frequent surface forms. This means surface span-F1 counts entity types, in contrast to entity tokens in the standard entity span-F1 metric.

As shown in Table 3, we first calculate the ratio of unique entities and total entities in each relevant train set (i.e., datasets where we do span labeling). A higher ratio number indicates a wider variety of spans. Both XLM-$R_{large}$ and ESCOXLM-R tend to have lower performance when variety gets high (above 0.75). In addition, there are 2 datasets (SAYFULLINA, JOBSTACK) where we see a low variety of spans and large discrepancy between performance of entity span-F1 and surface span-F1. This difference is lower for ESCOXLM-R (especially in SAYFULLINA) suggesting that pre-training on ESCO helps predicting uncommon entities.

It is also noteworthy that the standard deviations for the scores at the entity span-F1 are generally lower than those for the surface span-F1. This suggests that the results for the entity span-F1 scores are more consistent across different runs, likely due to recognizing common entities more.

Overall, ESCOXLM-R consistently outperforms XLM-$R_{large}$ in both the entity-level and surface-level F1 scores, indicating the benefits of using the ESCO dataset for pre-training on JAD tasks.

## 6 Related Work

To the best of our knowledge, we are the first to internalize an LM with ESCO for job-related NLP tasks. There are, however, several works that integrate factual knowledge (i.e., knowledge graphs/bases) into an LM. Peters et al. (2019) integrates multiple knowledge bases into LMs to enhance their representations with structured, human-curated knowledge and improve perplexity, fact recall and downstream performance on various tasks. Zhang et al. (2019); He et al. (2020); Wang et al. (2021b) combine LM training with knowledge graph embeddings. Wang et al. (2021a) introduces K-Adapter for injecting knowledge into pre-trained models that adds neural adapters for each kind of knowledge domain. Yu et al. (2022) introduces Dict-BERT, which incorporates definitions of rare or infrequent words into the input sequence and further pre-trains a BERT model.

Calixto et al. (2021) introduced a multilingual Wikipedia hyperlink prediction intermediate task to improve language model pre-training. Similarly, Yasunaga et al. (2022) introduced LinkBERT which leverages links between documents, such as hyperlinks, to capture dependencies and knowledge that span across documents by placing linked documents in the same context and pre-training the LM with MLM and document relation prediction.

## 7 Conclusion

In this study, we introduce ESCOXLM-R as a multilingual, domain-adapted LM that has been further pre-trained on the ESCO taxonomy. We evaluated ESCOXLM-R, to the best of our knowledge, on the broadest evaluation set in this domain on 4 different languages. The results showed that ESCOXLM-R outperformed XLM-R$_{\text{large}}$ on job-related downstream tasks in 6 out of 9 datasets, particularly when the task was relevant to the ESCO taxonomy and context was important. It was found that the improvement of ESCOXLM-R was mainly due to its performance on shorter span lengths, demonstrating the value of pre-training on the ESCO dataset. ESCOXLM-R also demonstrated improved performance on both frequent surface spans and a wider range of spans. Overall, this work showed the potential of ESCOXLM-R as an LM for multilingual job-related tasks. We hope that it will encourage further research in this area.

## Limitations

There are several limitations to this study that should be considered. First, a key limitation is the lack of a variety of language-specific JAD. Here, we have four different languages namely EN, DA, FR, and DE. This means that our analysis is based on a limited subset of languages and may not be representative of JAD data outside of these four languages.

In turn, the second limitation is that the ESCO taxonomy used as pre-training data only covers Europe and the datasets used in this work also covers mostly Europe. The results may not be generalizable to other regions. However, we see a slight improvement in the BHOLA dataset, the data of which comes from Singapore, which hints that it could generalize to other cultures.

The ESCO relation prediction task aims for learning the relations between elements of the ESCO taxonomy. We acknowledge that we do not evaluate the effectiveness of the pre-training objective in relation-centered tasks. Unfortunately, to the best of our knowledge, there is no job-related dataset containing relations between skill/occupation concepts to benchmark our model on. We consider this interesting future work.

Finally, we did not conduct an ablation study on the ERP pre-training objective, i.e., which errors it makes. As the accuracy of the objective is 60%, we are unable to determine which sampling method is detrimental to this accuracy. However, we suspect that the Linked sampling approach might be the hardest to predict correctly. For example, many occupations have a lot of necessary and optional skills, thus it is harder to determine if some skill truly belongs to a specific occupation. Nevertheless, we see that adding the ERP objective improves over regular MLM domain-adaptive pre-training.

Despite these limitations, we believe that this study provides valuable resources and insights into the use of ESCOXLM-R for analyzing JAD and suggests directions for future research. Future studies could address the limitations of this study by using a larger, more diverse datasets and by conducting ablation studies on the language model to better understand which parts contribute to the results.

## Ethics Statement

We also see a potential lack of language inclusiveness within our work, as we addressed in the Limitation section that ESCO mostly covers Europe (and the Arabic language). Nevertheless, we see ESCOXLM-R as a step towards inclusiveness, due to JAD frequently being English-only. In addition, to the best of our knowledge, ESCO itself is devoid of any gendered language, specifically, pronouns and other gender-specific terms in, e.g., occupations. However, we acknowledge that LMs such as ESCOXLM-R could potentially be exploited in the process of hiring candidates for a specific job with unintended consequences (unconscious bias and dual use). There exists active research on fairer recommender systems (e.g., bias mitigation) for human resources (e.g., Mujtaba and Mahapatra, 2019; Raghavan et al., 2020; Deshpande et al., 2020; Köchling and Wehner, 2020; Sánchez-Monedero et al., 2020; Wilson et al., 2021; van Els et al., 2022; Arafan et al., 2022).

## Acknowledgements

We thank both the NLPnorth and MaiNLP group for feedback on an earlier version of this paper. This research is supported by the Independent Research Fund Denmark (DFF) grant 9131-00019B and in parts by ERC Consolidator Grant DIALECT 101043235.

## References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Adam Mehdi Arafan, David Graus, Fernando P Santos, and Emma Beauxis-Aussalet. 2022. End-to-end bias mitigation in candidate recommender systems with fairness gates.

Krisztian Balog, Yi Fang, Maarten De Rijke, Pavel Serdyukov, and Luo Si. 2012. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2–3):127–256.

David Beauchemin, Julien Laumonier, Yvan Le Ster, and Marouane Yassine. 2022. "FIJO": a french insurance soft skill detection dataset. *arXiv e-prints*, pages arXiv–2204.

Akshay Bhola, Kishaloy Halder, Animesh Prasad, and Min-Yen Kan. 2020. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5832–5842, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Erik Brynjolfsson and Andrew McAfee. 2011. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Brynjolfsson and McAfee.

Erik Brynjolfsson and Andrew McAfee. 2014. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Marlis Buchmann, Helen Buchs, Felix Busch, Simon Clematide, Ann-Sophie Gnehm, and Jan Müller. 2022. Swiss job market monitor: A rich source of demand-side micro data of the labour market. *European Sociological Review*.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3651–3661.

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.

Mariia Chernova. 2020. Occupational skills extraction with FinBERT. *Master's Thesis*.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Design of negative sampling strategies for distantly supervised skill extraction. *ArXiv preprint*, abs/2209.05987.

Jens-Joris Decorte, Jeroen Van Hautte, Thomas Demeester, and Chris Develder. 2021. Jobbert: Understanding job titles through skills. *ArXiv preprint*, abs/2109.09605.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Ketki V Deshpande, Shimei Pan, and James R Foulds. 2020. Mitigating demographic bias in ai-based resume filtering. In *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization*, pages 268–275.

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.

ESCO. 2022. Machine Learning Assisted Mapping of Multilingual Occupational Data to ESCO (Part 1).

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.

Ann-Sophie Gnehm and Simon Clematide. 2020. Text zoning and classification for job advertisements in German, French and English. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93, Online. Association for Computational Linguistics.

Nidhi Goyal, Jushaan Kalra, Charu Sharma, Raghava Mutharaju, Niharika Sachdeva, and Ponnurangam Kumaraguru. 2023. JobXMLC: EXtreme multi-label classification of job skills with graph neural networks. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2181–2191, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Green, Diana Maynard, and Chenghua Lin. 2022. Development of a benchmark corpus to support entity recognition in job descriptions. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1201–1208, 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.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290, Online. Association for Computational Linguistics.

Sepp Hochreiter, Jürgen Schmidhuber, and Corso Elvezia. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Faizan Javed, Qinlong Luo, Matt McNair, Ferosh Jacob, Meng Zhao, and Tae Seung Kang. 2015. Carotene: A job title classification system for the online recruitment domain. In *2015 IEEE First International Conference on Big Data Computing Service and Applications*, pages 286–293. IEEE.

Faizan Javed, Matt McNair, Ferosh Jacob, and Meng Zhao. 2016. Towards a job title classification system. *ArXiv preprint*, abs/1606.00917.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 210–221, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Ilkka Kivimäki, Alexander Panchenko, Adrien Dessy, Dries Verdegem, Pascal Francq, Hugues Bersini, and Marco Saerens. 2013. A graph-based approach to skill extraction from text. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, pages 79–87, Seattle, Washington, USA. Association for Computational Linguistics.

Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Business Research*, 13(3):795–848.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Martin le Vrang, Agis Papantoniou, Erika Pauwels, Pieter Fannes, Dominique Vandensteen, and Johan De Smedt. 2014. Esco: Boosting job matching in europe with semantic interoperability. *Computer*, 47(10):57–64.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 115–124. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481.

Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 458–468.

Luiza Sayfullina, Eric Malmi, and Juho Kannala. 2018. Learning representations for soft skill matching. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 141–152.

Baoxu Shi, Jaewon Yang, Feng Guo, and Qi He. 2020. Salience and market-aware skill extraction for job targeting. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2871–2879. ACM.

Ellery Smith, Martin Braschler, Andreas Weiler, and Thomas Haberthuer. 2019. Syntax-based skill extractor for job advertisements. In *2019 6th Swiss Conference on Data Science (SDS)*, pages 80–81. IEEE.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Damian A Tamburri, Willem-Jan Van Den Heuvel, and Martin Garriga. 2020. Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 391–394. IEEE.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Sarah-Jane van Els, David Graus, and Emma Beauxis-Aussalet. 2022. Improving fairness assessments with synthetic data: a practical use case with a recommender system for human resources.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2022. Dict-BERT: Enhancing language model pre-training with dictionary. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1907–1918, Dublin, Ireland. 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.

Mike Zhang, Kristian Nørgaard Jensen, Rob van der Goot, and Barbara Plank. 2022c. Skill extraction from job postings using weak supervision. *ArXiv preprint*, abs/2209.08071.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Meng Zhao, Faizan Javed, Ferosh Jacob, and Matt McNair. 2015. SKILL: A system for skill identification and normalization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 4012–4018. AAAI Press.

## A Example Extraction from ESCO

```
{
    "id": int,
    "esco_code": "2250.4",
    "preferred_label": "animal therapist",
    "major_group": {
      "title": "Veterinarians",
      "description": "Veterinarians diagnose, [...]"
    },
    "alternative_label": [
      "animal convalescence therapist",
      "animal rehab therapist",
      "animal rehabilitation therapist",
      "animal therapists",
      "animal therapist"
    ],
    "description": "Animal therapists provide [...]",
    "essential_skills": [
      {
        "title": "anatomy of animals",
        "description": "The study of animal body parts, [...]"
      },
      ...
    ],
    "optional_skills": [
      {
        "title": "use physiotherapy for treatment of animals",
        "description": "Adapt human physical therapy [...]"
      },
      ...
  ]

}
```

Listing 1: **Example Extraction.** An example of the information that is given for ESCO code 2250.4: animal therapist. The original page can be found here: http://data.europa.eu/esco/occupation/0b2d3242-22a3-4de5-bd29-efd39cdf2c31.

```
Experience  O           O
in          O           O
working     B-Skill     O
on          I-Skill     O
a           I-Skill     O
cloud-based I-Skill     O
application I-Skill     O
running     O           O
on          O           O
Docker      O           B-Knowledge
.           O           O

A           O           O
degree      O           B-Knowledge
in          O           I-Knowledge
Computer    O           I-Knowledge
Science     O           I-Knowledge
or          O           O
related     O           O
fields      O           O
.           O           O
```

Listing 2: **Data Example SkillSpan.**

```
ability     O
to          O
work        B-Skill
under       I-Skill
stress      I-Skill
condition   O

due         O
to          O
the         O
dynamic     B-Skill
nature      O
of          O
the         O
group       O
environment O
,           O
the         O
ideal       O
candidate   O
will        O
```

Listing 3: **Data Example Sayfullina.**

## B Data Examples

| | |
|---|---|
| SKILLSPAN | Listing 2 |
| SAYFULLINA | Listing 3 |
| GREEN | Listing 4 |
| BHOLA | Listing 5 |
| KOMPETENCER | Listing 6 |
| FIJO | Listing 7 |
| GNEHM | Listing 8 |

Table 4: Data example references for each dataset.

```
A               O
sound           O
understanding   O
of              O
the             O
Care            B-Skill
Standards       I-Skill
together        O
with            O
a               O
Nursing         B-Qualification
qualification   I-Qualification
and             O
current         O
NMC             B-Qualification
registration    I-Qualification
are             O
essential       O
for             O
this            O
role            O
.               O
```

Listing 4: **Data Example Green.**

```
department economics national university singapore invites applications
teaching oriented positions level lecturer senior lecturer [...] <labels>
```

Listing 5: **Data Example Bhola.**

```
<English>
team worker                          S4
passion for developing your career   S1
liaise with internal teams           S1
identify system requirements         S2
plan out our new features            S4

<Danish>
arbejde med børn i alderen ½-3 år    S3
samarbejde                           S1
fokusere på god kommunikation        S1
bidrage til at styrke fællesskabet   S1
ansvarsbevidst                       A1
lyst til et aktivt udeliv            A1
```

Listing 6: **Data Example Kompetencer.**

```
Participer      B-relationnel
au              I-relationnel
réseau          I-relationnel
téléphonique    I-relationnel
mis             O
sur             O
pied            O
lors            O
des             O
campagnes       O
d'inscription   O
pour            O
fournir         B-pensee
les             I-pensee
renseignements  I-pensee
nécessaires     I-pensee
aux             I-pensee
assurés         I-pensee
```

Listing 7: **Data Example Fijo.**

```
in                  O
mit                 O
guten               O
EDV-Kenntnissen     B-ICT

.                   O
Es                  O
erwartet            O
Sie                 O
eine                O
interessante        O
Aufgabe             O
in                  O
einer               O
Adressverwaltung    O
(                   O
Rechenzenter        B-ICT
)                   O
```

Listing 8: **Data Example Gnehm.**

## C Fine-tuning Details

For fine-tuning XLM-R$_{\text{large}}$ (+ DAPT) and ESCOXLM-R on the downstream tasks, we use MaChAmp (van der Goot et al., 2021). For more details we refer to their paper. We always include the original learning rate, batch size, maximum sequence length, and epochs from the respective downstream tasks in our search space (whenever applicable). Each model is trained on an NVIDIA A100 GPU with 40GBs of VRAM and an AMD Epyc 7662 CPU. The seed numbers the models are initialized with are 276800, 381552, 497646, 624189, 884832. We run all models with the maximun number of epochs indicated in Table 5 and select the best-performing one based on validation set performance in the downstream metric.

| | Learning rate | Batch size | max_seq_length | Epochs |
|---|---|---|---|---|
| SKILLSPAN | {$1e^{-4}$, $5e^{-5}$, $1e^{-5}$ $5e^{-6}$} | {16, 32, 64} | 128 | 20 |
| KOMPETENCER | {$1e^{-4}$, $7e^{-5}$, $5e^{-5}$, $1e^{-5}$, $5e^{-6}$} | {8, 16, 32} | 128 | 20 |
| BHOLA | {$1e^{-4}$, $7e^{-5}$, $5e^{-5}$, $1e^{-5}$, $5e^{-6}$} | {4, 16, 32, 64, 128} | {128, 256} | 10 |
| SAYFULLINA | {$1e^{-4}$, $5e^{-5}$, $1e^{-5}$} | {16, 32, 64} | 128 | 10 |
| GREEN | {$1e^{-4}$, $5e^{-5}$, $1e^{-5}$} | {16, 32, 64} | 128 | 10 |
| JOBSTACK | {$1e^{-4}$, $7e^{-5}$, $5e^{-5}$, $1e^{-5}$, $5e^{-6}$} | {16, 32, 64, 128} | 128 | 20 |
| GNEHM | {$1e^{-4}$, $5e^{-5}$, $1e^{-5}$} | {16, 32, 64} | 128 | 5 |
| FIJO | {$1e^{-4}$, $5e^{-5}$, $1e^{-5}$} | {8, 16, 32, 64} | 128 | 10 |

Table 5: **Hyperparameter Sweep for Fine-tuning.** We show a hyperparameter sweep for fine-tuning all models. Learning rate differs for both XLM-$R_{large}$ and ESCOXLM-R, where XLM-$R_{large}$ performs best on lower learning rate (e.g., $1e^{-5}$) and ESCOXLM-R on a bit of a higher learning rate (e.g., $5e^{-5}$). A batch size of 32 works best for all models. The max sequence length is usually the same, except for BHOLA due to it containing long texts. Epochs are determined based on previous work (i.e., the relevant datasets).

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Limitations section*

☑ A2. Did you discuss any potential risks of your work?
*Ethics Statement section*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and Section 1*

☑ A4. Have you used AI writing assistants when working on this paper?
*Grammarly*

### B ☑ Did you use or create scientific artifacts?

*Section 1, Section 2, Section 3*

☑ B1. Did you cite the creators of artifacts you used?
*Section 1, 2, 3*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Footnote 1+2, Table 1, we plan to release the model under an Apache 2.0 License.*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Partially reflected in the Ethics Statement*

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*The JobStack dataset contains some privacy-bearing entities. We re-train the models on the anonymized dataset.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Table 1*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Table 1*

### C ☑ Did you run computational experiments?

*Section 2, 3, 4*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Section 2, and 4*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Appendix B*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 4, and 5*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section 2.2, and Appendix B*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*No response.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*No response.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*No response.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No response.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*No response.*