Cross-Lingual Sentence-Level Skill Identification in English and Danish Job Advertisements

Nurlan Musazade
Åbo Akademi University
Turku, Finland
nurlan.musazade@abo.fi

Mike Zhang
Aalborg University
Copenhagen, Denmark
jjz@cs.aau.dk

József Mezei
Åbo Akademi University
Turku, Finland
Jozsef.Mezei@abo.fi

Abstract

The increasing influence of artificial intelligence (AI), the availability of textual data, and large language models (LLMs) over the past decade is evident in the growth of scholarly work on identifying skills from job advertisements. In this work, we examine the detection of sentences that express skills as well as the explainability of model decisions with respect to their dependence on skill related tokens. We compare traditional machine learning (ML) approaches with a pretrained multilingual model and domain-adapted models for the task of English skill identification, and we assess the role of skill tokens in the classification process. We also investigate the ability of these models to generalize from English (EN) to Danish (DA) in both few-shot and zero-shot settings. Our findings indicate that both models achieve high performance in sentence classification achieving an F_1-score of 94% for EN and overall accuracy between 93%–94% for both EN and DA. The results show that traditional ML methods can remain relevant under certain circumstances reinforcing the importance of realistic baselines in the context of skill identification.

1 Introduction

With the technological advancements and labor market disruptions, the importance of identifying skill requirements in job advertisements rises for both job seekers and educational institutions (Brasse, 2024). Job advertisements are a critical source to study skill requirements (Khaouja et al., 2021; Senger et al., 2024; Zhang et al., 2022a). The most straightforward method is sentence-level skill identification (SI), where the task is to predict whether a sentence contains a skill or not (Khaouja et al., 2021).

Although SI methods have been studied for the English language (Tamburri et al., 2020; Khaouja et al., 2021; Leon et al., 2024a; Rosenberger, 2025), it is unclear how well this extends to other languages. We hypothesize, considering that some hard skills (e.g., Python, Java) are defined the same across languages, that there is a high generalizability level of English-based models. We extend prior research by testing the capabilities of (domain-adapted) multilingual language models (Chung et al., 2021; Zhang et al., 2023) on the task of SI, exploring the cross-lingual generalization of both a simple statistical baseline and fine-tuned models on English, and analyzing the factors that influence model decisions when classifying sentences. We seek to answer the following research questions:

RQ1 How effective are logistic regression, domain trained and pre-trained LLMs in classifying skill-related sentences in job advertisements?

RQ2 To what extent do English-based sentence classification models perform on Danish skill identification in zero- and few-shot settings?

RQ3 What is the contribution of skill tokens in the classification of skill-related sentences in job advertisements?

Contributions. In this work, we contribute the following by answering the RQs by showing that: (i) both the baseline and multilingual LMs show high performance on in-language (English) skill identification and cross-lingual skill identification for Danish, even with little training data; (ii) sentence-level SI models' reliance on skill tokens are low, highlighting the need to assess context-dependent trustworthiness and robustness of such approach in real-world scenarios.

2 Related Work

2.1 Classification-Based Skill Identification

Khaouja et al. (2021) identified four primary skill extraction techniques: skill counting, topic modeling, embeddings (skill or word-based), and ma-

Class	Train	Dev.	Test
0 - No skill	6,753	2,634	2,699
1 - Skill	10,421	3,359	3,056

Table 1: **Class Distribution.** Distribution of skill/no-skill sentences in the combined English dataset.

chine learning methods. From an ML perspective, there are two principal methods: Named Entity Recognition (NER), which classifies and extracts entities into predefined categories and content-based text classification (i.e., classifying whether a sentence contains a skill or not). We adopt the latter, applying sentence-level binary classification. Using sentence-level granularity in skill identification helps preserve word relationships and context.

Lin et al. (2023) and Rosenberger (2025) highlighted the critical role of removing irrelevant information for achieving high model performance. Lin et al. (2023) explored synthetic data generation to enhance data relevance. Rosenberger (2025) used classification approaches to filter noise from job ads, significantly improving recommendation accuracy. Their jobGBERT model achieved high accuracy (0.96-0.97 F_1 score) by truncating irrelevant content at the paragraph level. In contrast, our research addresses sentence-level classification, potentially enhancing model training simplicity and annotation efficiency. Additionally, while Rosenberger (2025) targeted the German language, our research investigates English and Danish datasets, exploring cross-lingual transfer and zero- or few-shot learning scenarios.

Leon et al. (2024b) conducted binary and multi-label classification using English and multilingual models on job ads. Facing imbalanced data with a prevalence of skill-absent sentences, the authors applied augmentation techniques. The accuracy of these models ranged from 94%–99%. They noted limitations in data availability and domain adaptation challenges and emphasized the need for further exploration of multilingual applicability and model explainability.

More recently, there has been increasing focus on computational efficiency in skill extraction from job advertisements (Sun et al., 2025; Vásquez-Rodríguez et al., 2024). For example, Vásquez-Rodríguez et al. (2024) compared various methods to distinguish their effectiveness and efficiency.

Our study builds upon these works, such as multilingual capabilities and explainability to improve real-time skill extraction and job recommendation systems. For a more detailed survey of SI, we refer to Khaouja et al. (2021).

Model	Class	F1	P	R	Acc.
LR + TF-IDF	0	0.89	0.91	0.87	0.90
	1	0.90	0.89	0.92	
RemBERT	0	0.93	0.96	0.90	0.93
	1	0.94	0.92	0.96	
ESCOXLM-R	0	0.93	0.95	0.91	**0.94**
	1	0.94	0.92	0.96	

Table 2: **English Results.** Performance of SI on the English test set in terms of F_1-score, precision (P), recall (R), and Accuracy (Acc.).

3 Methodology

Skill identification at the sentence level can be formulated as a binary classification task. Given a set of sentences $S = \{s_1, s_2, \ldots, s_n\}$, each sentence s_i is associated with a binary label y_i:

$$y_i = \begin{cases} 1 & \text{if sentence } s_i \text{ contains a skill mention,} \\ 0 & \text{otherwise.} \end{cases}$$

The goal of this task is to train a classification model f that accurately predicts the label y_i given a sentence s_i: $f(s_i) \rightarrow y_i$. The objective is to accurately classify each sentence on whether it contains a skill or not.

3.1 Data

We used three real-world English datasets for training and evaluation: SkillSpan (Zhang et al., 2022a), Green (Green et al., 2022), and Sayfullina (Sayfullina et al., 2018). Each dataset includes separate training, development, and test splits. In terms of dataset size, Green has 9,968, Sayfullina 7,411, and SkillSpan 11,543 sentences. In SkillSpan, there are both skill and knowledge annotations and were merged into one positive class. For Green, only positive skill annotations are taken.

Sentences were reconstructed from the word-level tokenized lists and each was labeled as either containing a skill (1) or not (0). Table 1 shows the distribution across all English splits. The combined data was shuffled before training. For Danish, we used the Kompetencer dataset (Zhang et al., 2022b), which follows the same format. It includes 778 training, 346 validation, and 262 test sentences.

Training	Class	RemBERT				ESCOXLM-R			
		F1	P	R	Acc.	F1	P	R	Acc.
DA	0	0.89	0.88	0.89	0.82	0.89	0.79	1.00	0.79
	1	0.55	0.57	0.54		0.00	0.00	0.00	
EN	0	0.95	0.98	0.92	0.92	0.95	0.97	0.92	0.92
	1	0.83	0.75	0.93		0.82	0.75	0.91	
EN + DA	0	0.95	0.96	0.95	0.93	0.95	0.96	0.95	0.93
	1	0.83	0.81	0.85		0.83	0.81	0.85	

Table 3: **Cross-lingual Results.** Cross-lingual performance on the Danish test set, broken down by training data, model and class

3.2 Models

A random baseline for the English test set yields around 0.51 accuracy. For a second baseline, we trained a TF-IDF-based logistic regression model with unigrams and bigrams (5,000 max features).

To conduct SI with language models, we experimented with three models in this study; namely RemBERT (Chung et al., 2021), ESCOXLM-R (Zhang et al., 2023) which is a domain-adapted XLM-R-based model (Conneau et al., 2020).

3.3 Training and Evaluation

For training the multilingual language models, we explored a limited range of hyperparameters and finalized on a learning rate of 1×10^{-6}, batch size of 16, weight decay on 0.01, and ten epochs with patience of 2. For both the English and few-shot Danish experiments, we kept these hyperparameters. We measured performance by F_1 score, precision, recall, and accuracy.

3.4 Explainability

For analysis, we investigate model explainability and compare model behavior. We use Integrated Gradients (IG; Sundararajan et al., 2017), implemented with the Captum library (Kokhlikyan et al., 2020), to analyze true positive (TP) predictions. The aim is to interpret which input tokens contribute most to skill predictions, improving transparency. IG satisfies two key axioms: Sensitivity, where a differing feature between inputs with different outputs must receive non-zero attribution; and Implementation Invariance, where models with identical outputs for all inputs yield the same attributions.

4 Results and Discussion

4.1 English Results

In Table 2, we show the main results for English of both the baselines and language models. We observe that all three models achieve a high F-score (0.89–0.94) as well as a high accuracy (0.90–0.94). In particular, from a cost–efficiency perspective, the baseline took a few minutes to train, whereas ESCOXLM-r and RemBERT took between 80–140 minutes, with equal performance.

Among the large language models, performance differences are minor but important, especially for correctly identifying skill-containing sentences. Both models perform equally well on the positive class. ESCOXLM-R achieves slightly higher overall accuracy and better recall for the negative class, while RemBERT shows marginally better precision on that class. Given the small differences and limited model tuning, no definitive conclusion can be drawn about one model being superior.

4.2 Danish Results

For our cross-lingual experiments, we have three setups

- **(DA)**: Fine-tuning the language models on the few-hundred Danish instances.
- **(EN)**: Fine-tuning the language models on English only and then apply it to the Danish test set.
- **(EN + DA)**: Fine-tuning both models on English and Danish and apply it to Danish.

In Table 3, we show that DA and EN demonstrate good performance; precision starts from 0.75 and minimum recall is 0.85 for the positive class. The ESCOXLM-R model failed to predicted no skill sentences with DA, whereas the RemBERT model's performance without the English language training is around 0.55 for the positive class.

Model	**Precision**				**Recall** set	**F1** set
	Top-1	Top-2	Top-3	Set		
LR + TF-IDF	0.485	0.278	0.255	0.225	0.808	0.303
ESCOXLM-R	0.101	0.172	0.213	0.234	0.576	0.284
RemBERT	0.098	0.170	0.209	0.231	0.576	0.281

Table 4: **Explainability.** Comparison of the matching True Skill with the tokens contributing positively for the TP

Text	**True Skills**	**Top Attributed Tokens**
javascript reactjs java	[javascript, reactjs, java]	[javascript, reactjs, java]
- and yaml	[yaml]	[-, ya]
Strong knowledge of application data and infrastructure architecture disciplines	[application, data, and, infrastructure, architecture]	[•, knowledge, application, data, infrastructure, s]
Demonstrated experience of performing DevOps for platforms	[DevOps, for, platforms]	[•ted, experience, of, performing, Dev, for]
You are proficient in Python and English	[Python, English]	[You, profi, in, English]

Table 5: **Dataset Sample.** Sample data rows.

Performance of the two models are similar in the other settings (i.e., EN and EN + DA). Interestingly, for the positive class, we see when the models have been fine-tuned on English data only (EN) it outperforms the DA setting (0.06–0.08 F_1 higher), likely due to more training data being available, indicating successful cross-lingual transfer.

4.3 Explainability

For analysis, we compare the contribution of actual skills to class prediction focusing on the TP class. Table 4 shows that the sets of words driving positive predictions are very similar across the two models and that ESCOXLM-R performance is only slightly higher. In the base model we multiply the TF-IDF values by the coefficients to measure each word's contribution to TP predictions and then compare that with word attribution in the LMs. Precision in the base model is higher for the top two tokens but it declines when we consider the full token set, ending up below the LMs. The F_1-score stays similar in the base model due to the high recall.

All models have a precision around 0.22–0.23 and a 0.58 recall for the LMs and 0.81 for the baseline. This suggests that classification does not depend mainly on semantic content or even on the explicit presence of a skill term in the sentence. These results align with the observation that ESCOXLM-R, despite its domain training, performs similarly to RemBERT and that logistic regression with TF-IDF narrows most of the gap with the LMs.

5 Conclusion

In this work, show the effectiveness of language models, including multilingual ones and a basic supervised ML model with TF-IDF, for the task of skill identification. All our models achieve around 90%-94% accuracy on English, with ESCOXLM-r as best-performing, indicating that the task is straightforward.

The performance of the supervised baseline model demonstrates that traditional approaches should still be considered after the introduction of the advanced architectures and models, and can be beneficial, for instance in the low resource settings. Furthermore, we show the effectiveness of multilingual LMs for cross-lingual transfer. We show that the performance of multilingual LMs is still high (92%), even though we only train on English data. As expected, there is overlap between skills between languages. This could particularly benefit low resource languages either in zero-shot setting or with the minimal training data.

In our explainability analysis, we show that the contribution of the skill tokens do not contribute that much to the actual correct prediction, warranting further investigation. We conducted analysis at the token level without cleaning data from special characters and stopwords, which poses a limitation in the evaluation of the explainability, while representing a realistic inputs (see Table 5). Computing attribution scores for each word, e.g., by summing tokens' attribution scores, may present a more reliable and interpretable definitive measurement.

For future research, we will consider more languages to investigate whether transfer still holds. Additionally, we considered all skills as one, but can also distinguish between hard and soft skills.

Ethics Statement

For the identification of specific occupational skills in sentences, we do not foresee any ethical issues.

References

Julia Brasse. 2024. Identification of future skills using data-driven methods: A systematic literature review and directions for future research. *Proceedings of the 57th Hawaii International Conference on System Sciences*.

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.

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

Imane Khaouja, Ismail Kassou, and Mounir Ghogho. 2021. A survey on skill identification from online job ads. *IEEE Access*, 9:118134–118153.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Florin Leon, Marius Gavrilescu, Sabina-Adriana Floria, and Alina Adriana Minea. 2024a. Hierarchical classification of transversal skills in job advertisements based on sentence embeddings. *Information*, 15(3).

Florin Leon, Marius Gavrilescu, Sabina-Adriana Floria, and Alina Adriana Minea. 2024b. Hierarchical classification of transversal skills in job advertisements based on sentence embeddings. *Information*, 15(3):151.

Shiyong Lin, Yiping Yuan, Carol Jin, and Yi Pan. 2023. Skill graph construction from semantic understanding. In *Companion Proceedings of the ACM Web Conference 2023*, pages 978–982.

Rosenberger. 2025. Careerbert: Matching resumes to esco jobs in a shared embedding space for generic job recommendations. *Expert Systems with Applications*.

Luiza Sayfullina, Eric Malmi, and Juho Kannala. 2018. Learning representations for soft skill matching. In *Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers 7*, pages 141–152. Springer.

Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 1–15, St. Julian's, Malta. Association for Computational Linguistics.

Ying Sun, Yang Ji, Hengshu Zhu, Fuzhen Zhuang, Qing He, and Hui Xiong. 2025. Market-aware long-term job skill recommendation with explainable deep reinforcement learning. *ACM Transactions on Information Systems*, 43(2):1–35.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

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.

Laura Vásquez-Rodríguez, Samuel Michel Bertrand Audrin, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa, and Lonneke van der Plas. 2024. Hardware-effective approaches for skill extraction in job offers and resumes. *RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems*.

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*

Thirteenth Language Resources and Evaluation Conference, pages 436–447, Marseille, France. European Language Resources Association.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2023. ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11871–11890, Toronto, Canada. Association for Computational Linguistics.