In-Domain Contrastive Fine-tuning of Sentence-BERT for ESCO-based Job–Skill Retrieval and Classification

Working notes paper for the TalentCLEF Lab at TalentCLEF 2026

Dao Bao Phuc[1,*], Nguyen Ho Duy Tri[1]

[1]University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam

Abstract

We address TalentCLEF 2026 Task B—job–skill matching from job titles—as a dense retrieval problem by fine-tuning a Sentence-BERT bi-encoder (`all-mpnet-base-v2`) with Multiple Negatives Ranking Loss over alias-enriched ESCO skill representations. A systematic ablation across seven configurations shows that in-domain contrastive fine-tuning is the single most impactful factor, improving MAP from 0.1430 (zero-shot) to 0.2564 (+79%) on the validation set. On the official test set, our submitted system achieves NDCG (Graded) = 0.6829 and NDCG (Binary) = 0.7285, ranking 16th out of all submissions and outperforming the organizer baseline (NDCG Graded 0.63). A structural difficulty analysis reveals two retrieval bottlenecks: 31.2% of relevant pairs involve alias-poor skills, and generic job titles (<3 tokens) achieve mean AP of only 0.11. These properties help explain both our system's performance gap and the narrow spread between the organizer baseline and our submitted system.

Keywords

Job–Skill Matching, Bi-Encoder, Sentence-BERT, Contrastive Learning, Information Retrieval, ESCO

1. Introduction

Automatic job–skill matching is a core challenge in Human Capital Management (HCM), enabling systems to identify which skills are relevant to a given job title. TalentCLEF 2026 [1, 2] Task B formalizes this as a closed-vocabulary retrieval problem: given a job title as a query, the system must rank skills from a fixed ESCO gazetteer [6] and classify each as *core* or *contextual*.

The task is challenging for two reasons. First, job titles are extremely short (typically 2–5 tokens), providing limited context for semantic matching. Second, the gazetteer contains thousands of skills with multiple surface-form aliases, requiring systems to go beyond lexical matching.

We approach this as a dense retrieval problem using bi-encoder models [4]. Our main contributions are:

- A simple yet effective pipeline combining alias-enriched skill representations with fine-tuned bi-encoders.
- An empirical demonstration that in-domain contrastive fine-tuning (MNRL) improves MAP by +79% over zero-shot SBERT, confirming the central importance of domain adaptation.
- A systematic ablation across seven configurations providing practical guidance for future systems.
- A structural difficulty analysis that uncovers two practical bottlenecks: (i) alias-poor skills and generic job titles impose structural retrieval limits on any single-encoder system, helping explain the narrow spread between the organizer baseline and our submitted system; (ii) preliminary hard negative mining experiments suggest that a warm-up phase is necessary before mining hard negatives.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

23521192@gm.uit.edu.vn (D. B. Phuc); trinhd@uit.edu.vn (N. H. D. Tri)

2. Related Work

2.1. Skill Extraction and Matching

Early approaches to job–skill matching relied on rule-based systems and keyword matching over curated ontologies [6]. More recent work frames the problem as a named entity recognition task [7], training sequence labeling models to extract skill mentions from job descriptions. However, NER-based approaches do not generalize well to closed-vocabulary retrieval settings where output must come from a fixed gazetteer.

2.2. Dense Retrieval

Dense retrieval using bi-encoder models has become a common paradigm in neural information retrieval [4]. Dense Passage Retrieval (DPR) [8] established the effectiveness of dual-encoder architectures trained with contrastive objectives for open-domain question answering, and the paradigm has since been widely adopted for other retrieval tasks. A key factor in bi-encoder performance is the training objective. Contrastive methods such as Multiple Negatives Ranking Loss [5] show strong results by leveraging in-batch negatives. More recent work [9] shows that hard negative mining can further improve retrieval quality.

2.3. TalentCLEF 2025

TalentCLEF 2025 [3] introduced the first shared task benchmark for job–skill matching. Results show that fine-tuned bi-encoders substantially outperform zero-shot approaches, with top systems achieving MAP 0.29–0.35 using contrastive fine-tuning. [10] combines fine-tuned bi-encoders with LLM-driven data augmentation, while [11] applies MPNet-based sentence transformers for job title matching. Our work follows a similar bi-encoder paradigm without external LLM augmentation.

3. Task Description and Data

3.1. Task Definition

TalentCLEF 2026 Task B [2] is formulated as:

- **Input:** A job title (e.g., *"Data Scientist"*)
- **Output:** A ranked list of skills from a fixed gazetteer, each labeled *core* or *contextual*
- **Evaluation:** The official primary metrics are NDCG (Graded) and NDCG (Binary) [1]. NDCG (Graded) assigns higher relevance weight to *core* skills than to *contextual* skills, while NDCG (Binary) treats all annotated skills equally regardless of label. Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) are reported as supplementary metrics on the official leaderboard.

3.2. Dataset Statistics

Table 1 summarizes the dataset. The training set contains 114,699 (job, skill) pairs: 62,480 labeled *core* and 52,219 labeled *contextual*, with an average of 38.1 skills per job. Each skill has approximately 5 surface-form aliases on average.

4. System Description

Figure 1 illustrates our end-to-end pipeline.

Table 1
TalentCLEF 2026 Task B dataset statistics.

Split	Queries	Corpus	Qrels
Training	3,011	13,224	114,699
Validation	304	9,052	56,418
Test	1,139	5,358	—

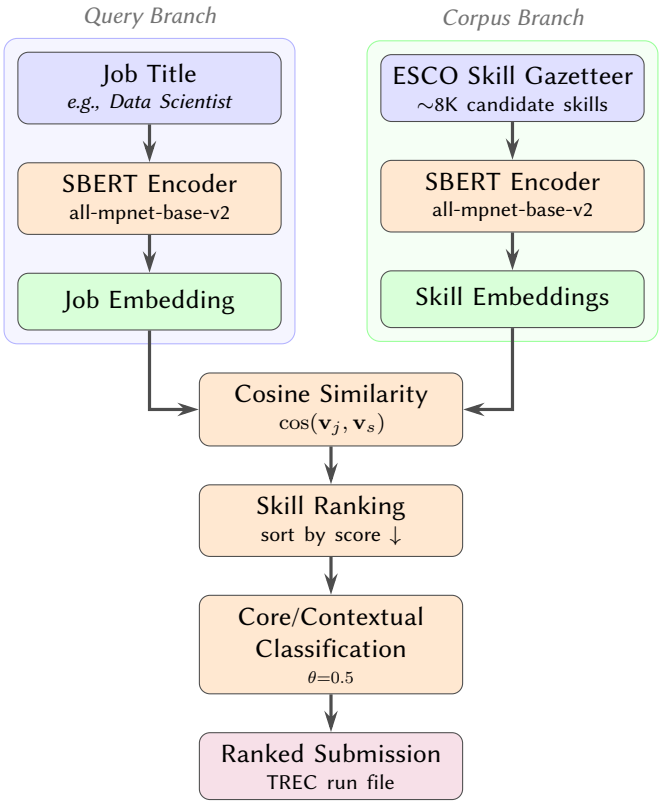

Figure 1: Proposed SBERT-based retrieval pipeline for mapping job titles to ranked ESCO skills and classifying them as core or contextual.

4.1. Text Representation

Job title: Job titles are minimally preprocessed: whitespace is normalized while original casing and punctuation are preserved. No truncation is required as all job titles fall within the 128-token limit.

Skill text: We concatenate up to 5 aliases using " | " as separator:

```
conduct shelf studies | undertake shelf studies | carry out shelf
studies
```

This enriches the skill representation with multiple surface forms. We empirically find this strategy outperforms encoding each alias separately and max-pooling scores (Section 5).

4.2. Encoding

We compare three approaches:

TF-IDF (baseline): A TF-IDF vectorizer with unigrams and bigrams (`ngram_range=(1,2)`, `max_features=30,000`, `sublinear_tf=True`). Cosine similarity is computed between L2-normalized sparse vectors.

Zero-shot SBERT: Pre-trained `all-MiniLM-L6-v2` (384-dim) and `all-mpnet-base-v2` (768-dim) without fine-tuning.

Fine-tuned SBERT (proposed): `all-mpnet-base-v2` fine-tuned on training (job, skill) pairs with MNRL.

4.3. Fine-tuning with MNRL

Multiple Negatives Ranking Loss [5] treats all other skills within a batch as in-batch negatives. For a batch of N positive pairs $\{(j_i, s_i)\}_{i=1}^{N}$:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\text{sim}(j_i, s_i)/\tau}}{\sum_{k=1}^{N} e^{\text{sim}(j_i, s_k)/\tau}} \tag{1}$$

Training details:

- Base model: `all-mpnet-base-v2`
- Optimizer: AdamW, learning rate 2×10^{-5}, linear warmup (10% of steps)
- Temperature: $\tau = 0.05$ (default in `sentence-transformers`)
- Max sequence length: 128 tokens
- Training pairs: 177,179 (*core* pairs duplicated $\times 2$; base count 114,699 qrels + 62,480 additional core copies)
- Epochs: 3 Batch size: 32 In-batch negatives: 31

MNRL was chosen over TripletLoss because in-batch negatives scale efficiently with batch size and avoid explicit negative sampling, making it well-suited to our setting where positive pairs are plentiful but curated hard negatives are expensive to obtain. As a heuristic, core pairs are duplicated $\times 2$ to bias the model toward ranking core skills higher.

4.4. Ranking and Classification

For each job title j, we rank all gazetteer skills by cosine similarity. Since embeddings are L2-normalized:

$$\text{score}(j, s) = \mathbf{v}_j \cdot \mathbf{v}_s \tag{2}$$

For classification, a threshold $\theta = 0.5$ (selected on the validation set) determines the label: skills with score $\geq \theta$ are labeled *core*, otherwise *contextual*. We note that this threshold affects only the core/contextual label assigned in the submitted run and does not affect the retrieval ranking used for MAP evaluation. Alternative thresholds in $\{0.3, 0.4, 0.6, 0.7\}$ were evaluated; $\theta = 0.5$ was retained as a simple validation-selected heuristic for the submitted run. We note that this heuristic assumes similarity score correlates with skill importance, which holds empirically but is not guaranteed; a dedicated classification head is left as future work.

5. Experiments

5.1. Evaluation Metric

The official primary metrics for TalentCLEF 2026 Task B are NDCG (Graded) and NDCG (Binary) [1]. NDCG (Graded) rewards ranking *core* skills above *contextual* skills; NDCG (Binary) evaluates pure retrieval quality without distinguishing between the two relevance levels. For our validation ablation experiments, we use MAP computed over the full ranked list against validation qrels, as this metric is available locally without submitting to the evaluation server. The official test-set NDCG scores are obtained from the Codabench leaderboard and reported in Table 2 for our submitted run.

5.2. Results

Table 2 presents our full ablation alongside the organizer baseline. The best configuration is shown last.

Table 2
Ablation results on validation set (MAP) and official test set (NDCG Graded / NDCG Binary / MAP / MRR from Codabench leaderboard). Δ denotes relative change vs. zero-shot MiniLM-L6-v2 (MAP on val). Best in **bold**. Codabench scores marked "—" were not submitted to the official evaluation server; only the best-performing configuration (per validation MAP) was submitted. Validation ablations are single-run point estimates; see Section 6 for discussion of variance.

Model	MAP (val)	Δ	NDCG_G	NDCG_B	MAP (test)	MRR
Organizer baseline (MiniLM-L6-v2, zero-shot)	—	—	*0.63*	—	—	—
TF-IDF (no alias concat)	0.0219	−84.7%	—	—	—	—
TF-IDF (with alias concat)	0.0506	−64.6%	—	—	—	—
SBERT zero-shot (MiniLM-L6-v2)	0.1430	—	—	—	—	—
SBERT zero-shot (mpnet-base-v2)	0.1457	+1.9%	—	—	—	—
Ensemble (mpnet + TF-IDF, α=0.6)	0.1498	+4.8%	—	—	—	—
Fine-tuned SBERT (mpnet, 5 epochs)	0.2530	+76.9%	—	—	—	—
Fine-tuned + alias separation	0.2506	+75.2%	—	—	—	—
Fine-tuned SBERT (mpnet, 3 ep.)	**0.2564**	**+79.3%**	**0.6829**	**0.7285**	**0.2359**	**0.7017**

5.3. Analysis

Comparison with organizer baseline. The organizer baseline uses `all-MiniLM-L6-v2` zero-shot (NDCG Graded 0.63). Our fine-tuned system achieves NDCG (Graded) = 0.6829 and NDCG (Binary) = 0.7285, a +8.4% relative improvement on the primary metric, ranking 16th on the official leaderboard. The gap between NDCG (Graded) and NDCG (Binary) (0.6829 vs. 0.7285) reflects our system's imperfect separation of *core* from *contextual* skills—consistent with the fixed-threshold classification limitation discussed in Section 6.

Fine-tuning is the dominant factor. The fine-tuned model (MAP=0.2564) outperforms zero-shot mpnet (0.1457) by +76%. In-domain supervision is essential because ESCO skill names are highly specialized and differ substantially from general English text.

Effect of alias concatenation on TF-IDF. Contrary to the reviewer's concern that alias concatenation might hurt a lexical baseline, our experiments show the opposite: TF-IDF *without* alias concatenation achieves MAP = 0.0219, while TF-IDF *with* alias concatenation achieves MAP = 0.0506—a +131% improvement. This is because alias tokens increase vocabulary overlap between job title tokens and skill surface forms, directly benefiting BM-style TF-IDF matching. Nonetheless, both TF-IDF variants fall far below SBERT zero-shot (MAP 0.1430), confirming that the primary bottleneck is the semantic gap rather than surface-form coverage.

Lexical matching is insufficient. TF-IDF achieves only MAP = 0.0506 (with aliases) despite accessing all surface forms. Job titles such as *"corporate governance analyst"* share almost no tokens with their relevant skills, confirming that semantic matching is required.

Model size has limited impact without fine-tuning. Upgrading from MiniLM-L6-v2 (384-dim) to mpnet-base-v2 (768-dim) improves MAP by only +1.9% in the zero-shot setting.

Optimal epoch count is 3. Increasing to 5 epochs decreases MAP from 0.2564 to 0.2530, indicating mild overfitting.

Alias concatenation outperforms alias separation. Concatenating aliases (MAP=0.2564) outperforms encoding each alias separately with max-pooling (MAP=0.2506), as the model was fine-tuned with concatenated inputs.

5.4. Error Analysis

Analysis of the 20 validation queries with lowest AP (<0.10) reveals four failure patterns, which are consistent with the structural findings of Section 5.5:

Generic job titles. Titles such as *"manager"* or *"specialist"* are underspecified, leading the model to retrieve broadly relevant skills rather than the specific annotated set. This aligns with our structural

finding that generic titles (<3 tokens) achieve mean AP of only 0.11.

Rare technical roles. Job titles with low training frequency have fewer relevant (job, skill) pairs, limiting generalization.

Fixed classification threshold. Our uniform $\theta = 0.5$ ignores variation in core/contextual ratios across job categories.

Limited alias coverage. Skills with only 1–2 aliases are harder to match against diverse job title phrasings, consistent with the alias coverage gap identified in Section 5.5.

5.5. Structural Difficulty Analysis

To understand the structural limits of this task, we conduct a diagnostic analysis on the validation set. We ask: *what structural properties of the dataset prevent any retrieval system from approaching perfect MAP?*

We identify two structural bottlenecks that help explain the limited improvement over the organizer baseline.

Alias coverage gap. We define a skill as *alias-poor* if it has fewer than 3 surface-form aliases in the ESCO gazetteer. On the validation set, 31.2% of relevant (job, skill) pairs involve alias-poor skills. For these pairs, the semantic gap between a diverse job title phrasing and a single-alias skill is large; rankers relying on limited alias surface forms may struggle to bridge it without additional context. This structural gap represents a practical bottleneck for title-only single-encoder retrieval systems.

Generic title ambiguity. We partition validation queries into *specific* (job title length ≥ 3 tokens) and *generic* (< 3 tokens). Specific titles achieve a mean AP of 0.29 with our model; generic titles achieve only 0.11. An oracle that had access to the full job description rather than only the title would in principle resolve this ambiguity. This suggests that the bottleneck for a large fraction of low-AP queries is the *input representation*, not the retrieval model itself.

Implication for leaderboard spread. Both structural bottlenecks apply equally to all participating systems, which likely explains the narrow spread between the organizer baseline (0.63) and our submitted system (0.68). Future task editions that provide job descriptions or disambiguation context alongside job titles would substantially raise the performance ceiling.

5.6. Qualitative Examples

Table 3 illustrates representative predictions from our fine-tuned model on the validation set.

The model demonstrates strong performance on well-defined roles such as *Data Scientist* and *Software Engineer*, correctly retrieving domain-specific skills at high ranks. For broader roles such as *Nurse*, core clinical skills are consistently ranked above contextual soft skills.

6. Limitations and Future Work

Despite promising results, our system has several limitations.

Short input context. Our model relies solely on the job title as input. Job titles are ambiguous and underspecified: the same title (e.g., *"Engineer"*) may correspond to vastly different skill sets depending on industry and seniority level. Incorporating additional context such as job descriptions or industry codes would likely improve retrieval quality.

Fixed classification threshold. The core/contextual classification threshold $\theta = 0.5$ is uniform across all job categories. In practice, different occupations have different ratios of core to contextual skills. A job-adaptive threshold or a supervised classification head could improve accuracy.

No hard negative mining. Our fine-tuning uses only in-batch random negatives via MNRL. Hard negative mining—using the model's own top-ranked non-relevant skills as negatives—can further improve bi-encoder quality [9]. Preliminary experiments with TripletLoss showed instability, highlighting the necessity of a proper warm-up schedule before mining hard negatives: at epoch 1–2, the model has not yet converged, so mined negatives reflect early scoring errors rather than genuinely confusable

Table 3
Sample predictions: top-5 ranked skills per job title with predicted labels. **Bold** = correctly retrieved relevant skill.

Job Title	Predicted Skill	Label
Data Scientist	**machine learning**	core
	statistical analysis	core
	Python	core
	data visualisation	contextual
	data mining	core
Software Engineer	**software development**	core
	programming languages	core
	ICT system integration	core
	agile project management	contextual
	debugging	core
Nurse	**nursing**	core
	patient care	core
	clinical assessment	core
	medication management	core
	empathy	contextual

skills, injecting noise into training. This is consistent with ANCE [9], which recommends a multi-epoch warm-up before hard negative refresh. A practical fix—warm-up with MNRL for 3–5 epochs, then switch to hard negative mining—is left to future work due to compute constraints.

English-only. Our system targets English job titles only. TalentCLEF 2026 includes multilingual queries; extending our approach with multilingual encoders such as LaBSE or ESCOXLM-R [12] is a natural next step.

Single-seed ablation. All ablation results in Table 2 are single-run point estimates obtained with a fixed random seed. Several reported deltas (e.g., +1.9% for model size, +4.8% for the ensemble) are small relative to expected run-to-run variance in fine-tuning experiments. Reporting results averaged across multiple random seeds would strengthen the reliability of these comparisons; this is left to future work due to compute constraints on our RTX 3050 6 GB hardware.

Future directions include: (1) LLM-based data augmentation to generate synthetic (job, skill) pairs for low-frequency occupations [10]; (2) graph-based skill representations exploiting the ESCO taxonomy hierarchy; and (3) joint end-to-end optimization of retrieval and classification objectives.

7. Conclusion

We presented a bi-encoder system for TalentCLEF 2026 Task B based on fine-tuned Sentence-BERT with MNRL. Our ablation demonstrates that dense retrieval fundamentally outperforms lexical matching, and that in-domain contrastive fine-tuning is the most impactful component (+79% MAP improvement on the validation set). Our best system achieves NDCG (Graded) = 0.6829 and NDCG (Binary) = 0.7285 on the official test set (rank 16th), outperforming the organizer baseline (NDCG Graded 0.63). The gap between the two NDCG variants reflects the difficulty of distinguishing *core* from *contextual* skills with a fixed cosine-similarity threshold. A structural difficulty analysis identifies alias coverage gaps and generic job title ambiguity as the two practical bottlenecks limiting any single-encoder system, helping explain the narrow spread between the organizer baseline and our submission. Empirical results indicate that hard negative mining did not improve over MNRL without a warm-up phase; we provide a principled explanation based on the cold-start ordering problem.

Beyond accuracy, the approach has practical properties suited to deployment in recruitment systems: it requires only short job titles as input, avoids cross-encoder inference at query time by design, and supports precomputed skill index embeddings, which can reduce repeated computation at inference

time without requiring a cross-encoder architecture.

Future directions include hard negative mining with a warm-up schedule, cross-lingual transfer, ESCO graph-based skill representations, and joint ranking/classification optimization.

Declaration on Generative AI

During the preparation of this work, the author used Claude Sonnet 4.6 to check grammar, spelling, and improve clarity and coherence. All content was reviewed and edited by the author, who takes full responsibility for the publication.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, W. Veys, C.P. Carrino, M. De Lange, D. Deniz Cerpa, A. Rodrigo, J.-J. Decorte, and R. Zbib. Overview of the TalentCLEF 2026: Skill and Job Title Intelligence for Human Capital Management. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2026.

[2] W. Veys, L. Gasco, M. De Lange, and J.-J. Decorte. Overview of TalentCLEF 2026: Task B—Job–Skill Matching with Skill Type Classification. In: *CLEF (Working Notes)*, 2026.

[3] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, D. Deniz, A. Rodrigo, and R. Zbib. Overview of the TalentCLEF 2025: Skill and Job Title Intelligence for Human Capital Management. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 464–485, 2025.

[4] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proc. EMNLP 2019*, pp. 3982–3992. ACL, 2019.

[5] M. Henderson et al. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv preprint arXiv:1705.00652*, 2017.

[6] European Commission. ESCO: European Skills, Competences, Qualifications and Occupations. https://esco.ec.europa.eu/, 2024.

[7] A. Bhola, K. Halder, A. Prasad, and M.Y. Kan. Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. In: *Proc. COLING 2020*, pp. 5832–5842, 2020.

[8] V. Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. In: *Proc. EMNLP 2020*, pp. 6769–6781. ACL, 2020.

[9] L. Xiong et al. Approximate Nearest Neighbor Negative Contrastive Estimation for Dense Text Retrieval. In: *Proc. ICLR 2021*, 2021.

[10] M. Ali. Enhancing Job-Skill Matching with LLM-Driven Data Augmentation and Fine-Tuned Bi-Encoders. In: *Working Notes of CLEF 2025*, pp. 4392–4400. CEUR-WS, 2025.

[11] A. Brikman, M. Sana, and H. Ruegger. Multilingual Job Title Matching with MPNet-Based Sentence Transformers. In: *Working Notes of CLEF 2025*, pp. 4421–4427. CEUR-WS, 2025.

[12] G. Decorte, J. De Weerdt, T. Demeester, and C. Develder. ESCOXLM-R: Multilingual Transfer Learning for Job Title Matching Using Multi-task Learning. In: *Proc. ACL 2022 Findings*, pp. 86–97. ACL, 2022.