Hybrid Dense-Sparse Retrieval and Re-Ranking for Multilingual Resume Retrieval from Job Descriptions*

Working Notes for TalentCLEF 2026 Task A

Alonso Palomino[1,*]

[1]*Independent Researcher*

Abstract

This paper describes a submission to TalentCLEF 2026 Task A, the contextualized job-person matching task. The objective is to rank candidate resumes for a given job description in a multilingual setting. The study evaluates a two-stage retrieval architecture based on JobBERT as the first-stage dense retriever and compares two re-ranking strategies: a sparse lexical re-ranker based on SPLADE and a late-interaction re-ranker based on multilingual ColBERT. On the official development split, the strongest overall system is obtained with JobBERT followed by SPLADE-based reciprocal-rank-fusion re-ranking. ColBERT remains competitive in Spanish and provides a more interaction-rich alternative for long-form resume matching. The best dense-sparse system is used as the primary run.

Keywords

TalentCLEF, job-person matching, resume retrieval, dense retrieval, SPLADE, ColBERT, multilingual information retrieval

1. Introduction

TalentCLEF 2026 Task A focuses on contextualized job-person matching, where each query is a full job description and the objective is to retrieve and rank the most relevant candidate resumes [1, 2]. This setting is particularly challenging because job descriptions and resumes are long, structured, multilingual, and information-dense. As a result, effective systems need to capture both broad semantic compatibility and precise lexical evidence such as skills, tools, certifications, and role-specific terms.

Viewed more broadly, retrieval over skill requirements and competence descriptions extends beyond job-to-candidate matching. Closely related problems also arise in educational settings, where systems must identify assessment items that measure targeted vocational skills. Prior work has shown that semantic retrieval and re-ranking can support vocational exam assembly by retrieving relevant test items from validated item banks and by improving ranking quality under practical constraints such as bias control [3, 4]. This broader perspective highlights that skill-aware retrieval methods are useful wherever textual evidence must be matched to explicit competency requirements.

For this reason, a two-stage architecture is adopted. In the first stage, JobBERT is used as a dense retriever to generate a high-quality candidate set. In the second stage, the candidates are re-ranked using either a sparse lexical signal from SPLADE or a late-interaction signal from multilingual ColBERT. This design makes it possible to study two complementary ways of strengthening the first-stage semantic ranking.

The paper addresses the following research question: can a two-stage multilingual retrieval pipeline improve resume retrieval from job descriptions over a strong JobBERT baseline by adding a second-stage re-ranker with either sparse lexical matching or late interaction? The working hypothesis is that second-stage re-ranking will improve retrieval effectiveness over JobBERT alone, and that sparse re-ranking with SPLADE will be more robust overall than zero-shot multilingual ColBERT because explicit lexical evidence such as skills, tools, and credentials remains central to resume relevance.

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

*Search runs and code are available at https://github.com/alonsopg/talentclef2026-taskA.

*Corresponding author. Email: mail@alonsopg.com.

mail@alonsopg.com (A. Palomino)

The main finding is that re-ranking helps, but the choice of second-stage model matters: SPLADE is the most robust overall, while multilingual ColBERT remains competitive, especially in Spanish.

2. Task Description

TalentCLEF 2026 Task A evaluates contextualized job-person matching in a multilingual retrieval setting [2]. Given a job description, the system must rank candidate resumes according to their relevance to the job requirements. The submission focuses on the monolingual `en-en` and `es-es` tracks.

The official submission format follows the TREC run-file convention, with one run file per language pair. Submissions are evaluated with standard information retrieval metrics, with Mean Average Precision (MAP) as the main ranking criterion and MRR and nDCG reported as additional metrics [5, 6].

3. Data and Preprocessing

3.1. Official Data

This work uses the official TalentCLEF 2026 Task A development and test sets [7]. The development data contains 10 queries and 472 resumes for each of English and Spanish. The test data contains 40 queries and 476 resumes for each of the two languages.

3.2. Resume Text Construction

The final systems do not use the earlier ESCO-based or knowledge-graph-based re-ranking components explored in preliminary experiments. Instead, a text representation is built for each resume by extracting and concatenating the main resume sections, including profile, experience, skills, education, certifications, and languages. Both resumes and job descriptions are normalized before retrieval.

This preprocessing design keeps the retrieval pipeline simple and reproducible while preserving the information most likely to matter for candidate ranking. For reproducibility, whitespace was normalized after section concatenation and no translation, external supervision, ESCO expansion, or demographic metadata was used in the final submitted runs.

4. Method

4.1. First-Stage Dense Retrieval with JobBERT

As the first-stage retriever, this submission uses `TechWolf/JobBERT-v3`, a dense encoder from the JobBERT family developed for labor-market matching tasks [8]. JobBERT encodes both job descriptions and resume representations into a shared embedding space. Candidate resumes are then ranked by similarity to the job-description embedding.

Conceptually, JobBERT is used here as a domain-adapted bi-encoder. The job description and each resume are encoded independently into dense vectors, which makes retrieval efficient because resume embeddings can be computed in advance and compared to the query embedding with a vector-similarity operation. Its main strength is broad semantic matching when compatible roles or skills are expressed with different wording.

4.2. Sparse Candidate Re-Ranking with SPLADE

To complement dense retrieval with stronger lexical matching, JobBERT candidates are re-ranked using SPLADE [9, 10]. SPLADE produces sparse lexical representations that retain neural expressiveness while remaining sensitive to exact or near-exact term overlap.

Unlike a dense bi-encoder, SPLADE maps text into a high-dimensional sparse vector over the vocabulary. Important terms receive larger weights, and the model can also activate related expansion terms

learned by the transformer. This makes SPLADE useful for re-ranking in settings where exact evidence such as software names, certifications, or specific skills can strongly influence relevance.

SPLADE is applied over the top 100 JobBERT candidates and the two ranking signals are combined through reciprocal-rank fusion over the candidate set. The fusion uses `k=60`, with weight 2.0 for the JobBERT rank and weight 1.0 for the SPLADE rank. This yields the strongest overall system on development.

4.3. Late-Interaction Re-Ranking with ColBERT

As a more interaction-rich alternative, the study evaluates a pretrained multilingual ColBERT re-ranker using `jinaai/jina-colbert-v2` [11, 12]. Unlike bi-encoder retrieval, ColBERT preserves token-level interactions between query and document representations and is therefore appealing for long-form resume ranking, where local skill- and role-level matches may be important.

More specifically, ColBERT encodes query tokens and document tokens separately and delays most of the interaction until scoring time. For each query token, the model finds the best matching document token and then aggregates those token-level matches into a final relevance score. This late-interaction design is more expensive than single-vector retrieval, but can capture fine-grained alignments important in resume ranking.

ColBERT is applied as a second-stage re-ranker over the top 100 JobBERT candidates. The implementation uses query length 128, document length 220, query batch size 4, document batch size 2, and CPU inference. This variant is not the strongest overall system, but it is highly competitive in Spanish and provides a stronger innovation story than a purely dense-sparse hybrid.

5. Experimental Setup

5.1. Systems Compared

The following systems are compared on the official development split:

- **BM25**: lexical baseline.
- **SPLADE**: sparse neural baseline.
- **JobBERT**: dense first-stage retrieval baseline.
- **JobBERT→SPLADE**: candidate re-ranking with sparse lexical evidence.
- **JobBERT→ColBERT**: candidate re-ranking with late interaction.

Augmentation and compact-query variants were also explored in preliminary experiments, but these did not outperform the main re-ranking systems and are therefore treated as secondary ablations.

5.2. Implementation

JobBERT retrieval is implemented with SentenceTransformers [13]. PyTerrier is used for sparse retrieval experiments, and the SPLADE re-ranker is implemented through the `pyt_splade` stack. The late-interaction re-ranker is implemented with a pretrained multilingual ColBERT model through the PyLate stack. All official submissions output the top 100 documents per query in TREC format.

6. Results

6.1. Development Results

Table 1 summarizes the main development results used for model selection.

The results show that JobBERT is already a strong dense baseline. Re-ranking with SPLADE yields the best overall MAP in both English and Spanish. The ColBERT re-ranker does not surpass the SPLADE re-ranker overall, but in Spanish it is nearly tied in MAP and slightly stronger in nDCG. This makes ColBERT an attractive additional variant for analysis and submission.

Table 1
Main development results on TalentCLEF 2026 Task A.

Language	System	MAP@50	nDCG@50
en	JobBERT	0.7744	0.9245
en	JobBERT→SPLADE_RRF	0.7808	0.9294
en	JobBERT→ColBERT	0.7728	0.9225
es	JobBERT	0.7299	0.8962
es	JobBERT→SPLADE_RRF	0.7350	0.9034
es	JobBERT→ColBERT	0.7348	0.9086

6.2. Official Test Submission

Table 2 summarizes the official test-set runs submitted for TalentCLEF 2026 Task A. The primary run is the dense-sparse system selected from the development experiments. The mixed run keeps the best English dense-sparse variant and uses the late-interaction ColBERT variant for Spanish. The ColBERT run evaluates the late-interaction approach in both languages. The primary dense-sparse run obtains the best official average MAP among the three submitted runs.

Table 2
Official TalentCLEF 2026 Task A test-set submissions.

Run	Avg. MAP	MAP en	MAP es	MRR en	MRR es	nDCG en	nDCG es
Primary	0.5055	0.5505	0.4606	0.8074	0.7744	0.7323	0.6808
Mixed	0.4897	0.5505	0.4289	0.8074	0.7565	0.7323	0.6605
ColBERT	0.4812	0.5336	0.4289	0.8420	0.7565	0.7336	0.6605

7. Analysis

The experiments suggest that the strongest gains come from combining dense semantic retrieval with a complementary re-ranking signal. SPLADE improves on the dense retriever by reinforcing exact lexical evidence, which remains highly relevant in resume matching. This is especially useful when important job requirements are expressed through specific skills, technologies, tools, or credentials.

ColBERT offers a different advantage. By preserving token-level interactions, it is better aligned with the long-document nature of job descriptions and resumes than a plain single-vector model. Although it was not the best system overall on development, its strong Spanish performance suggests that late interaction is a viable direction for future work in multilingual job-person matching.

Several additional variants did not help. Standalone sparse retrieval was clearly weaker than JobBERT, long query augmentations were often neutral because they were truncated by the encoder, and compact rewrite variants were not robust, especially in Spanish. These negative results suggest that extra complexity is useful only when the added signal is both informative and compatible with the retrieval model's input limits.

8. Limitations

This work has several limitations. The public development set is small, which makes model selection fragile; the public data does not expose gender labels, which limits direct fairness analysis; and ColBERT was not fine-tuned because public supervision is limited and zero-shot re-ranking was already computationally demanding.

9. Conclusion

A two-stage retrieval framework for TalentCLEF 2026 Task A was presented based on JobBERT first-stage retrieval and alternative re-ranking strategies. The best development results were obtained with a hybrid dense-sparse re-ranking approach using SPLADE, while multilingual ColBERT provided a competitive and more distinctive late-interaction alternative, especially in Spanish. On the official test set, the primary dense-sparse run achieved an average MAP of 0.5055 across English and Spanish. These findings highlight the value of combining domain-adapted dense retrieval with complementary re-ranking signals for contextualized multilingual job-person matching. In relation to the research question, the results indicate that second-stage re-ranking can improve over a strong JobBERT baseline in multilingual resume retrieval from job descriptions. The working hypothesis is therefore only partially confirmed: re-ranking was beneficial overall, and SPLADE was the most robust approach, while zero-shot multilingual ColBERT remained competitive but did not consistently surpass the dense-sparse hybrid. Overall, strong semantic retrieval remains the foundation of this task, and its best results come from adding a complementary lexical or late-interaction signal rather than from increasing complexity indiscriminately.

Declaration on Generative AI

During the preparation of this work, the author used OpenAI GPT-5.4/Codex in order to help structure the manuscript, convert a paper outline into a LaTeX draft, and revise wording. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

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, 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] H. Fabregat, L. García-Sardiña, P. Estrella, L. Gasco, C. P. Carrino, D. Deniz Cerpa, A. Rodrigo, R. Zbib, Overview of talentclef 2026: Task a — contextualized job-person matching, in: CLEF (Working Notes), 2026.

[3] A. Palomino, A. Fischer, J. Kuzilek, J. Nitsch, N. Pinkwart, B. Paassen, EdTec-QBuilder: A semantic retrieval tool for assembling vocational training exams in German language, in: K.-W. Chang, A. Lee, N. Rajani (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 26–35. URL: https://aclanthology.org/2024.naacl-demo.3/. doi:10.18653/v1/2024.naacl-demo.3.

[4] A. Palomino, A. Fischer, D. Buschhüter, R. Roller, N. Pinkwart, B. Paassen, Mitigating bias in item retrieval for enhancing exam assembly in vocational education services, in: W. Chen, Y. Yang, M. Kachuee, X.-Y. Fu (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 183–193. URL: https://aclanthology.org/2025.naacl-industry.16/. doi:10.18653/v1/2025.naacl-industry.16.

[5] TalentCLEF Organisers, TalentCLEF 2026 Evaluation Info, https://talentclef.github.io/talentclef/docs/talentclef-2026/evaluation/evaluation_info/, 2026. Accessed: 2026-04-27.

[6] TalentCLEF Organisers, TalentCLEF 2026 Submission Format, https://talentclef.github.io/talentclef/docs/talentclef-2026/evaluation/submission_format/, 2026. Accessed: 2026-04-27.

[7] TalentCLEF Organisers, TalentCLEF 2026 Task, https://talentclef.github.io/talentclef/docs/, 2026. Accessed: 2026-04-27.

[8] L. Ribeiro, P. Loo, et al., Jobbert: Understanding job titles through skills, arXiv preprint arXiv:2109.09605 (2021). URL: https://arxiv.org/abs/2109.09605.

[9] T. Formal, B. Piwowarski, S. Clinchant, Splade: Sparse lexical and expansion model for first stage ranking, arXiv preprint arXiv:2107.05720 (2021). URL: https://arxiv.org/abs/2107.05720.

[10] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, Splade v2: Sparse lexical and expansion model for information retrieval, arXiv preprint arXiv:2109.10086 (2021). URL: https://arxiv.org/abs/2109.10086.

[11] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over BERT, in: Proceedings of SIGIR 2020, 2020. URL: https://arxiv.org/abs/2004.12832.

[12] R. Jha, B. Wang, M. Gunther, G. Mastrapas, S. Sturua, I. Mohr, A. Koukounas, V. Plachouras, J. D. Simonsen, Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever, in: Proceedings of the 1st Workshop on Multilingual Representation Learning, 2024. URL: https://aclanthology.org/2024.mrl-1.11/.

[13] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese BERT-networks, in: Proceedings of EMNLP-IJCNLP 2019, 2019. URL: https://aclanthology.org/D19-1410/.