Multilingual Job Title Matching with MPNet-Based Sentence Transformers*

Adam Brikman[1],*, Michael Sana[1] and Holden Ruegger[1]

[1] *Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332*

Abstract

We compare a pretrained multilingual sentence transformer to a fine-tuned variant for the TalentCLEF 2025 competition, which focuses on retrieving semantically similar job titles in a target language given a source-language query. Our baseline model is MPNet, a transformer with 278 million parameters. Fine-tuning was performed using Multiple Negatives Ranking Loss (MNRL) on in-domain monolingual job title pairs. The resulting model achieved a mean average precision (MAP) score of 0.360, placing 32nd on the public leaderboard. All code is publicly available at https://github.com/dsgt-kaggle-clef/talentclef-2025.

Keywords

NLP, Human Capital Management, Multilinguality, Cross-lingual Capability, Job Title Ranking, MPNet, CEUR-WS

1. Introduction

The inaugural Multilingual Job Title Matching task was held in 2025 and was part of the TalentCLEF lab [1] as part of the Conference and Labs of the Evaluation Forum (CLEF). This task seeks to select a job title and subsequently locate and rank similar job titles both from a monolingual and cross-lingual perspective. Evaluations from a monolingual perspective include English (EN), Spanish (ES), and German (DE) while cross-lingual are between English to Spanish (EN-ES) and English to German (EN-DE). The training dataset contains 28,880 English, 20,724 Spanish, and 23,023 German pairs of job titles, which were leveraged to fine-tune our MPNet-based model for monolingual job matching using a contrastive loss function.

We empirically demonstrate that fine-tuning a multilingual sentence transformer on finite job title pairs can significantly degrade cross-lingual performance due to domain overfitting and catastrophic forgetting. Furthermore, we have found that the pretrained MPNet model maintains strong cross-lingual alignment and outperforms the fine-tuned model. Our results suggest that leveraging a pretrained model may be preferable in similar tasks where training data is limited.

2. Overview

To address the task of multilingual job title retrieval, we rely on MPNet, a transformer-based sentence encoder known for its ability to generate semantically meaningful embeddings across languages [2, 3]. Our system encodes each job title into a dense vector representation, enabling efficient similarity-based retrieval across both monolingual and cross-lingual settings. Below, we describe the MPNet model and embedding process, followed by our rationale for using both pretrained and fine-tuned variants tailored to specific subtasks.

2.1. Sentence Transformer Model

MPNet (Masked and Permuted Pre-training Network) is a multilingual transformer encoder that builds on BERT and XLNet, integrating both masked language modeling and permuted language modeling

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

abrikman3@gatech.edu (A. Brikman); msana3@gatech.edu (M. Sana); hruegger3@gatech.edu (H. Ruegger)

0009-0006-6428-2264 (A. Brikman); 0009-0007-6608-7347 (M. Sana); 0009-0004-7811-8954 (H. Ruegger)

objectives [2, 4]. This hybrid pretraining strategy allows MPNet to capture deep semantic relationships across different languages, making it well suited for paraphrase mining and multilingual retrieval tasks.

Each job title is tokenized using MPNet's internal word-piece tokenizer, which segments text into subword units. A [CLS] token is prepended to the input and is used by the model to generate a global sentence-level embedding. This [CLS] vector, with a dimensionality of 768, is extracted from the final hidden layer and serves as the job title's semantic representation.

The model was pretrained using contrastive learning on a large-scale multilingual corpus. In this setting, positive sentence pairs are pulled closer in the embedding space, while negative pairs are pushed apart. This training objective helps preserve alignment across languages, enabling zero-shot and few-shot generalization in cross-lingual retrieval scenarios.

2.2. Model Variants

As outlined in Figure 1, our final system incorporates two MPNet variants: a pretrained model, and a fine-tuned version adapted to the job title dataset. The fine-tuning process updates all parameters, not just the classification head, using a contrastive loss objective on provided English, Spanish, and German job title pairs [3, 5].

Each variant was deployed strategically: the pretrained model was used for cross-lingual retrieval tasks (EN to ES and EN to DE), where maintaining robust multilingual alignment was critical. The fine-tuned model was reserved for monolingual subtasks (EN to EN, ES to ES, DE to DE), where domain-specific adaptation was expected to yield performance gains.

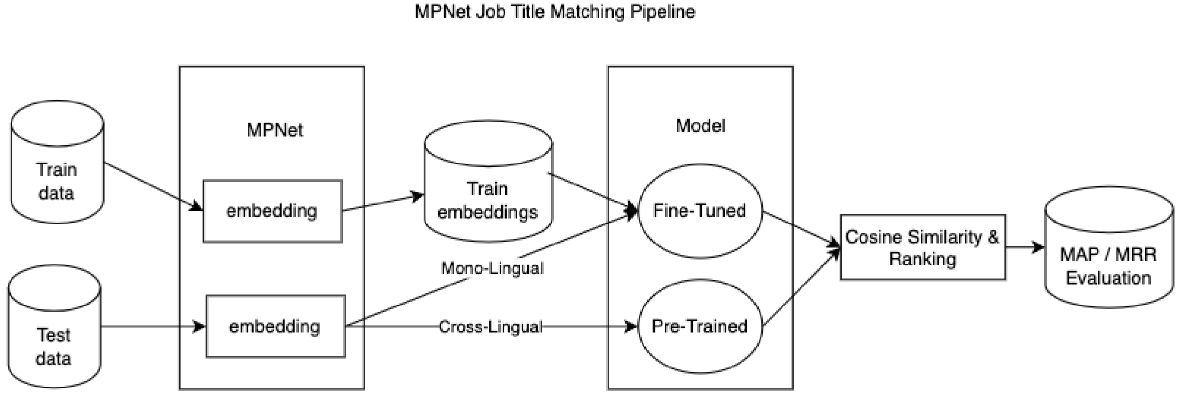

Figure 1: Transfer learning pipeline for multilingual job title retrieval. MPNet generates 768-dimensional embeddings from both train and test data. Fine-tuned MPNet is trained using monolingual job title pairs with Multiple Negatives Ranking Loss (MNRL), while the pretrained model is used directly for cross-lingual tasks. Cosine similarity is used to rank job titles, and performance is evaluated via MAP and MRR.

3. Methodology

Our implementation leveraged a lightweight pipeline utilizing pretrained multilingual sentence embeddings for cross-lingual job title retrieval and a fine-tuned model for monolingual retrieval [5]. Training was limited to the available positive job title pairs on the English, German, and Spanish training sets, while evaluation was performed on the validation sets. The following subsections outline our data handling, fine-tuning, training strategy, and evaluation methodology.

3.1. Data Handling and Representation

During training, we extracted positive sentence pairs from the provided TSV files in the English, Spanish, and German training sets. Each file consisted of European Skills, Competences, Qualifications and

Occupations (ESCO) job ID numbers, associated URLs, and job title pairs representing semantically similar occupations [1, 6, 7].

Job titles were provided in a consistently formatted manner, free of punctuation and in all lower-case characters. As a result, preprocessing such as stopword filtering, punctuation removal, or lowercasing was not implemented. All text was passed directly into the sentence transformer model, which natively handled tokenization, special tokens, and padding for varying job title length.

3.2. Model Architecture

We used the Multilingual MPNet Base V2 model as the foundation for both our pretrained and fine-tuned variants [3]. The transformer model contains 278 million parameters and produces 768-dimensional embeddings, based on the sentence transformers implementation built on top of MPNet [2]. We selected MPNet Base V2 for its strong cross-lingual performance, earning scores of approximately 0.84 for EN-ES and EN-DE in both Pearson and Spearman correlations on sentence similarity benchmarks [8]. Both metrics are commonplace in evaluating semantic alignment, evaluating how well a model captures meaning relationships between languages [8]. We selected this model for its strong performance on cross-lingual tasks and accessibility through the sentence transformer framework.

To address both monolingual and cross-lingual retrieval scenarios, we used two variants of the model. The pretrained variant was used for cross-lingual tasks (EN-ES and EN-DE) while the fine-tuned variant was optimized for monolingual retrieval (EN-EN, ES-ES, and DE-DE) by training on positive job title pairs from the English, Spanish, and German training sets.

While fine-tuning improved monolingual retrieval on the validation set, we observed impaired cross-lingual performance, which we attribute to overfitting and loss of multilingual generalization due to the limited linguistic and contextual diversity of the training set [9, 10].

3.3. Training Strategy

We fine-tuned the MPNet Base V2 model using a Multiple Negatives Ranking Loss (MNRL) function, which is a contrastive objective that treats all other samples in the batch as implicit negatives [5]. The MNRL function is suitable for tasks involving only positive pairs as it assumes all other pairs in the batch are unrelated to the anchor. This strategy is effective for Task A since it does not require negative cases and only positive pairs are given in the training set [1, 7].

The model was trained over 5 epochs with a batch size of 32. We used a learning rate of 1e-6, epsilon of 2e-5, and a weight decay of 0.01. These hyperparameters were selected to minimize the risk of overfitting and catastrophic forgetting [10]. Warmup steps were set to 10% of the total training steps and the best-performing model (based on training loss) was automatically saved.

No random seed was set during training, so minor variations may occur between runs. As will be noted in the future work section, reproducibility can be improved by establishing a random seed at the beginning of the implementation.

3.4. Evaluation Metrics

Our model's performance was evaluated on the provided validation set using Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), which were the primary metrics used in the leaderboard for this task [1]. Job title embeddings were compared using cosine similarity and titles were ranked based on their similarity to each query embedding.

Evaluation was performed on monolingual (EN-EN, ES-ES, DE-DE, ZH-ZH) and cross-lingual (EN-ES, EN-DE) scenarios. Chinese (ZH-ZH) was included in the evaluation despite the absence of training data, which allowed for the assessment of the fine-tuned model's zero-shot multilingual generalization.

4. Results

This section presents our model's performance across monolingual and cross-lingual retrieval tasks, as evaluated on the TalentCLEF test set. As seen in Table 1, our model achieved an average MAP of 0.360 on monolingual retrieval tasks, ranking 32nd on the public leaderboard. Performance was strongest on EN-EN (0.408), while lower scores were observed on ES-ES (0.348) and DE-DE (0.324). Notably, the model achieved a MAP of 0.380 on the ZH-ZH test set despite no fine-tuning on Chinese data, suggesting a level of zero-shot generalization being preserved from multilingual pretraining.

Table 1
Table 1 highlights a comparison between our model's test set monolingual MAP scores and the top two submissions on the public leaderboard. Our model ranked 32nd and the results underscore the challenges of fine-tuning multilingual transformers on finite domain-specific data. The performance gap supports our findings on the trade-offs between domain adaptation and cross-lingual generalization.

Team name	Rank	Avg. MAP	MAP(EN-EN)	MAP(ES-ES)	MAP(DE-DE)	MAP(ZH-ZH)
ranabarakat	1	0.530	0.559	0.527	0.516	0.508
omokhtar	2	0.554	0.522	0.515	0.507	0.516
DS@GT-TalentCLEF	32	0.360	0.408	0.348	0.324	0.380

Table 2 summarizes our model's performance on the cross-lingual objectives (EN-ES and EN-DE). Despite utilizing the pretrained multilingual MPNet model, we observed near-zero MAP scores (0.023 for EN-ES and 0.019 for EN-DE). Upon further review, we identified a language mismatch in the inference pipeline: for both EN-ES and EN-DE evaluations, job titles were mistakenly embedded in English, rather than in Spanish or German, as required. This resulted in invalid similarity comparisons between queries and unrelated job embeddings, leading to a collapse in retrieval performance. Future work will correct this by ensuring proper language-specific embedding and alignment during cross-lingual evaluation.

Table 2
Table 2 outlines a comparison between our model's cross-lingual test set MAP scores and the top two submissions on the public leaderboard. Note that ranabarakat's first place submission did not include cross-lingual scores. As such, the top two cross-lingual submissions earned second and third rank.

Team name	Rank	MAP(EN-ES)	MAP(EN-DE)
omokhtar	2	0.516	0.512
ranabarakat	3	0.516	0.516
DS@GT-TalentCLEF	32	0.023	0.019

These findings highlight the importance of aligning evaluation pipelines with model expectations, particularly for cross-lingual objectives. While monolingual retrieval remained reasonably effective, the collapse in cross-lingual performance illustrates the fragility of multilingual generalization when inputs are misaligned or when training data lacks sufficient linguistic diversity. These trade-offs form the basis for further reflection in our Discussion section.

5. Discussion

Our findings challenge the conventional wisdom that task-specific fine-tuning improves retrieval performance. Despite training on in-domain monolingual data, the fine-tuned MPNet model consistently underperformed the pretrained model on both monolingual and cross-lingual objectives. Fine-tuning appears to have disrupted the model's multilingual alignment and impaired generalization. Notably for cross-lingual tasks, during test evaluation, we identified an inference-time bug that caused all job title embeddings, regardless of language, to be generated in English. This mismatch led to invalid query-target comparisons and a collapse in similarity scores for EN-ES and EN-DE. As illustrated in

Figure 2, this degradation in cross-lingual semantic alignment is visually apparent. As a result, these cross-lingual results should be interpreted with caution.

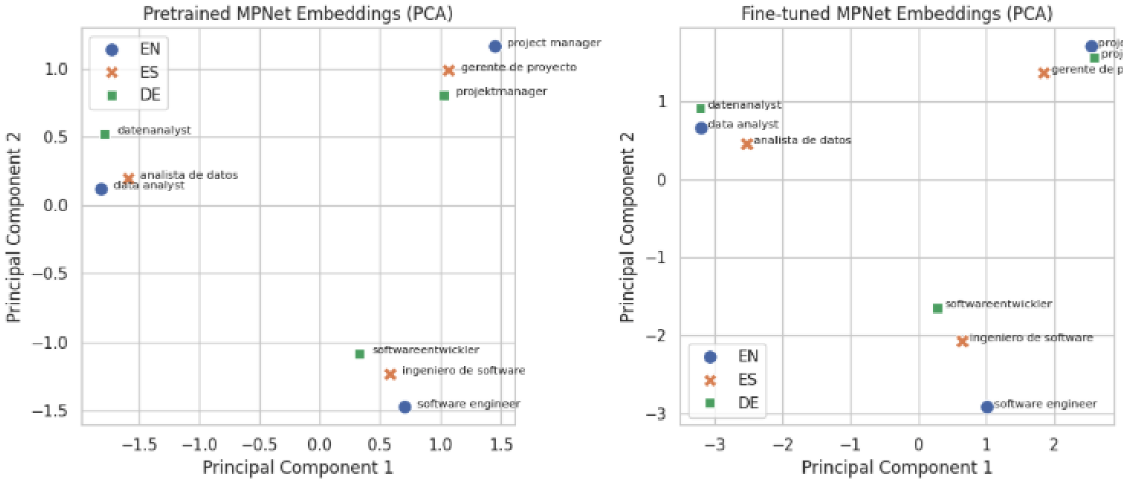

Figure 2: PCA visualization of job title embeddings before (left) and after (right) fine-tuning. The pretrained MPNet model maintains cross-lingual alignment, with semantically equivalent job titles clustering together across English (EN), Spanish (ES), and German (DE). After fine-tuning on monolingual data, this alignment deteriorates, demonstrating a shift in the embedding space that undermines cross-lingual retrieval performance.

Despite this, the pretrained model demonstrated promising robustness in the ZH-ZH setting, where it had no prior exposure to Chinese. This underscores the strength of large-scale multilingual pretraining for zero-shot generalization and raises important questions about when and how fine-tuning should be applied in multilingual contexts.

Fine-tuning on limited, single-language data introduced harmful overfitting while failing to deliver expected improvements, even on languages seen during training. In contrast, the unmodified pretrained model proved more consistent and robust across both seen and unseen languages. Prior work has shown that catastrophic forgetting [9], distortion of pretrained features [10], and the erosion of multilingual alignment during fine-tuning [5] are common risks when adapting pretrained models. Techniques such as language-specific regularization, contrastive alignment objectives [5], or parameter-efficient tuning strategies like adapter modules [11] have been proposed to mitigate these effects and preserve the benefits of multilingual pretraining.

Our findings echo concerns raised in recent work [5, 9, 10] that, without explicit safeguards, fine-tuning multilingual transformers on narrow monolingual datasets can degrade performance by distorting the pretrained embedding space, even in-language.

6. Future Work

Future work should include a broader evaluation of multilingual models beyond MPNet to determine whether alternative architectures offer improved retrieval performance. One such candidate is ESCOXLM-R, a model specifically trained on job market data and designed for multilingual representation learning [12]. To enhance reproducibility, the fine-tuning process should also be repeated with a fixed random seed—an element omitted here due to time constraints. Further research should explore bias-controlled training approaches that reduce performance disparities across gendered job titles. Finally, re-generating the job title embeddings as outlined in the preceding section may offer a straightforward path to performance improvement.

7. Conclusion

In this work, we fine-tuned a multilingual MPNet transformer encoder to retrieve semantically similar job titles in a target language, given a source-language query. However, the fine-tuned model underperformed relative to the pretrained baseline, yielding lower MAP scores across all evaluated language pairs. The final submission achieved an average MAP score of 0.360, placing 32nd on the public leaderboard of the TalentCLEF 2025 competition. Future improvements could involve exploring pretrained models specifically developed for the human resources domain, as well as incorporating enhanced language-specific embeddings for job titles. Our code is publicly available at https://github.com/dsgt-kaggle-clef/talentclef-2025

Acknowledgments

We want to thank the Data Science at Georgia Tech (DS@GT)-CLEF group for cloud infrastructure and their support, and the organizers of TalentCLEF for hosting the competition.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4) by OpenAI to assist with grammar, spelling, and phrasing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, D. Deniz, A. Rodrigo, 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, Springer, 2025.

[2] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, 2020. URL: https://arxiv.org/abs/2004.09297. arXiv:2004.09297.

[3] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, arXiv preprint arXiv:1906.08237 (2019).

[5] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. URL: https://arxiv.org/abs/2004.09813. arXiv:2004.09813.

[6] European Commission, ESCO - European Skills, Competences, Qualifications and Occupations, 2024. URL: https://esco.ec.europa.eu/en, accessed: 2024-05-25.

[7] L. Gascó, H. Fabregat, L. García-Sardiña, D. D. Cerpa, P. Estrella, Á. Rodrigo, R. Zbib, Talentclef 2025 corpus: Skill and job title intelligence for human capital management, 2025. URL: https://doi.org/10.5281/zenodo.15292308. doi:10.5281/zenodo.15292308.

[8] AIDA-UPM, mstsb-paraphrase-multilingual-mpnet-base-v2, https://huggingface.co/AIDA-UPM/mstsb-paraphrase-multilingual-mpnet-base-v2, 2024. Accessed: 2024-05-25.

[9] S. Kotha, J. M. Springer, A. Raghunathan, Understanding catastrophic forgetting in language models via implicit inference, 2024. URL: https://arxiv.org/abs/2309.10105. arXiv:2309.10105.

[10] A. Kumar, A. Raghunathan, R. Jones, T. Ma, P. Liang, Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. URL: https://arxiv.org/abs/2202.10054. arXiv:2202.10054.

[11] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, 2019. URL: https://arxiv.org/abs/1902.00751. arXiv:1902.00751.

[12] M. Zhang, R. van der Goot, B. Plank, ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain, 2023. URL: https://aclanthology.org/2023.acl-long.662.