Beyond Titles: Semantic Matching of Jobs and Skills Using LLMs and S-BERT*

Notebook for the AI Research Group in the ITCL Technology Center at CLEF 2025

Iago X. Vázquez[1,*], Rodrigo Sedano[1], Silvia González[1] and Javier Sedano[1]

[1]ITCL Technology Center, 70 López Bravo St., 09001 Burgos, Spain

Abstract

This paper presents an overview of the participation in Task B of the edition of TalentCLEF, a shared task held within the Conference and Labs of the Evaluation Forum (CLEF) 2025. Task B focuses on the challenge of matching job titles to relevant professional skills, a core problem in human capital management. Our team developed a system based on a novel ensemble approach that integrates representations from multiple large language models, which are combined and refined using a Sentence-BERT (S-BERT) model. We describe our methodology, datasets, and evaluation results, showing notable improvements over the official baseline, which uses job and skill titles alone. Our best system achieved a MAP of 0.2442, clearly surpassing the baseline MAP of 0.1874.

Keywords

Human Resources Management, Skill Extraction, Sentence Embeddings, Semantic Retrieval

1. Introduction

The TalentCLEF [1] initiative aims to advance research in natural language processing (NLP) applied to human resources management (HRM). In its 2025 edition [1], TalentCLEF introduced Task B, focused on developing systems that can automatically retrieve relevant professional skills given a job title. This capability is critical in improving recruitment, workforce planning, and career recommendation systems.

This paper presents the system developed by the AI Research Group in the ITCL Technology Center for the TalentCLEF 2025 challenge. The proposed solution is an ensemble method that integrates three distinct configurations of large language models (LLMs) [2], each involving different combinations of prompts and models. These configurations generate semantically rich textual representations of job titles and skills, which are subsequently encoded into dense embeddings using the S-BERT-based [3] model `all-MiniLM-L6-v2` [2]. Finally, cosine similarity is used to compare the resulting embeddings, and similarities across the different configurations are aggregated. We report the performance of each individual configuration, as well as the overall ensemble, based on standard information retrieval metrics.

The overview of Task B of TalentCLEF 2025 can be found in [4].

2. Task Description and Data Sources

Task B of TalentCLEF 2025 consists of a job-skill matching challenge. To develop our model, we used two datasets provided by the TalentCLEF 2025 organization team. These datasets are the Development

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

*Corresponding author.

iago.vazquez@itcl.es (I. X. Vázquez); rodrigo.sedano@itcl.es (R. Sedano); silvia.gonzalez@itcl.es (S. González); javier.sedano@itcl.es (J. Sedano)

0000-0002-5438-7656 (I. X. Vázquez); 0009-0000-7891-5751 (R. Sedano); 0000-0003-2095-3338 (S. González); 0000-0002-4191-8438 (J. Sedano)

[1]https://talentclef.github.io/talentclef/

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Set, intended for experimenting with different approaches, and the Test Set, used for the final evaluation of the developed algorithms.

The list of skills used in this task was obtained from the European Skills, Competences, Qualifications and Occupations (ESCO) [3] database. ESCO is a tool of the European Union that categorizes and links skills, competences, qualifications, and occupations to improve employability, education, and labour mobility across Europe. It facilitates interoperability between labour market and education systems.

Each provided dataset includes two files that define the model inputs:

- *queries*: This file contains a list of occupation names, each identified by a unique key.
- *corpus_elements*: This file contains a list of skills extracted from the ESCO database. Each skill is identified by a unique key. The corresponding URIs, along with various alternate titles available in ESCO, are included.

The results obtained from both datasets are evaluated against a ground truth, assessing the ranked relevance of skills in relation to each occupation.

3. Evaluation

The evaluation of the models focused on ranking all available skills from a provided source according to their relevance to each job title. The used evaluation metrics are:

- *Mean Average Precision (MAP)*: Measures the average precision at multiple levels of retrieval for each query, then averages the results across all queries. This metric is particularly suitable for evaluating information retrieval systems, as it accounts for the ranking order of the retrieved results.
- *Mean Reciprocal Rank (MRR)*: Computes the average of the reciprocal ranks of the first relevant result for each query. It is particularly effective in tasks where prioritizing the early retrieval of at least one relevant result is critical.
- *Normalized Discounted Cumulative Gain (nDCG)*: Evaluates the quality of a ranking by considering both the relevance and position of relevant documents, penalizing those that appear lower in the list. It is normalized to allow comparison across queries.
- *Precision at different cutoffs (P@X):* Measures the proportion of relevant documents among the top X results. It provides a meaningful assessment of ranking performance, especially in contexts where the highest-ranked outputs are of primary importance to users.

4. Baseline

The official baseline provided for Task B of TalentCLEF 2025 was adopted. This baseline computed the cosine similarity between the embeddings of occupation titles and skill names, as provided in the files. Since multiple alternative titles are available for each skill, the highest similarity score between any of these and the job title was taken as the skill's similarity value. The embeddings were generated using the `all-MiniLM-L6-v2` S-BERT model, which was also adopted in our approach.

Although this method establishes a useful reference point, it relies exclusively on surface-level lexical similarities. Therefore, it may overlook deeper semantic connections that could arise when contextual cues, or definitions, are incorporated.

5. Proposal Description

5.1. Overview

The approach for the accomplishment of Task B is based on computing semantic similarity between job and skill definitions. The pipeline consists of the following steps:

[3] https://esco.ec.europa.eu/es

- *Definition Generation*: Each job title and skill is defined using prompt-based queries to three different configurations, involving combinations of two distinct LLMs,two different prompts for job titles and two distinct prompts for skills. These combinations, denoted as $C1$, $C2$ and $C3$, are described in detail in Subsection 5.2. As a result, three distinct definitions are generated for each job title and each skill. Although multiple alternate titles were available for each skill, the first one was selected in each case to generate the definition, as it corresponds to the preferred term in ESCO.
- *Sentence Embedding*: All the generated definitions are encoded using the `all-MiniLM-L6-v2` S-BERT model, denoted as E.
- *Similarity Scoring*: For each job title and configuration C_i, the cosine similarity is computed between the job definition and each skill definition generated under the same configuration.
- *Ensemble Aggregation*: For each job title, the similarity scores of each skill across the three configurations are averaged to produce a final ranking. The skills are then ordered based on these aggregated scores. In this way, the relative importance of each skill for a given job is determined.

In Fig. 1, a schema representing the pipeline used to generate similarity between a job title and a skill is presented. In the figure, it can be seen how the cosine similarities between the job titles and the skills are merged to form the ensemble aggregation.

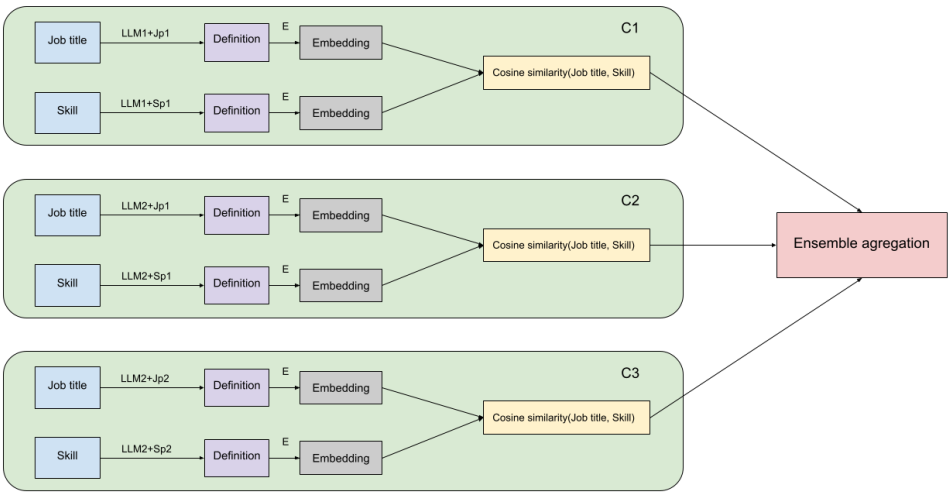

Figure 1: Pipeline overview for computing semantic similarity between a job title and a skill. Each configuration (C_1, C_2, C_3) employs a distinct combination of LLM, job title prompts (Jp_i), and skill prompts (Sp_i) to generate definitions for both, the input job title and the input skill. These definitions are encoded using the `all-MiniLM-L6-v2` (E) model, and compared via cosine similarity. The final similarity score for each skill is obtained by aggregating the similarity scores from the three configurations through ensemble averaging.

5.2. Combinations of LLMs and Prompts

To generate the definitions, we used three different configurations, based on two distinct large language models (LLMs), two prompts for job titles and two prompts for skills. For each configuration, a single query was executed per occupation or skill.

The two LLMs employed are:

- *Model 1 (LLM_1)*: `gemma-3-4b-it-qat` [5]
- *Model 2 (LLM_2)*: `gemma3:27b` [5]

LLM_1 is a quantized version of LLM_2, with 4 billion parameters, while LLM_2 is the original unquantized version, with 27 billion parameters. Both models were used with a temperature setting of 0.7 and the following system role prompt: 'You are an expert in the labour market. Answer in English.'

The prompts used for job titles are:

- *Job Prompt (Jp_1)*: 'Please define, outlining its functions, in less than fifty words, the following occupation: <OCCUPATION>. Write the definition in a similar way to the following example: "The techniques, theories, and commonly accepted strategies regarding pricing of goods. The relation between pricing strategies and outcomes in the market such as profitability maximisation, deterrence of newcomers, or increase of market share.". Make sure you only write the definition.'
- *Job Prompt (Jp_2)*: 'Please define, outlining its functions, in less than fifty words, the following occupation: <OCCUPATION>.'

The prompts used for skills are:

- *Skill Prompt (Sp_1)*: 'Please define, outlining its utility, in less than fifty words,the following skill: <SKILL>. Write the definition in a similar way to the following example: "The techniques, theories, and commonly accepted strategies regarding pricing of goods. The relation between pricing strategies and outcomes in the market such as profitability maximisation, deterrence of newcomers, or increase of market share.". Make sure you only write the definition.'
- *Skill Prompt (Sp_2)*: 'Please define, outlining its utility, in less than fifty words,the following skill: <SKILL>.'

In these prompts, the symbols '<' and '>' indicate the position where the occupation or skill name was inserted.

Finally, the combinations of models and prompts produced the following configurations:

- *Configuration 1 (C_1)*: $LLM_1 + Jp_1 + Sp_1$
- *Configuration 2 (C_2)*: $LLM_2 + Jp_1 + Sp_1$
- *Configuration 3 (C_3)*: $LLM_2 + Jp_2 + Sp_2$

Configurations C_1 and C_2 used the same prompts (Jp_1 and Sp_1) but different models, whereas C_3 employed alternative, simpler prompts (Jp_2 and Sp_2). In contrast, C_1 used the quantized model LLM_1, while C_2 and C_3 used the unquantized one LLM_2.

In Fig. 2, an example of the generated definitions is shown. There, unique keys identifying skills from the Development Set are associated with their corresponding definitions, obtained through the configuration C_1, which is composed of the LLM `gemma-3-4b-it-qat`, along with the job prompt Jp_1 and the skill prompt Sp_1.

```
dev_cb_sk_1 Pricing plans are structured offerings outlining product/service costs and features. They're crucial for revenue generation, customer acquisition, and competitive positioning within a market, directly impacting sales and overall business strategy.
dev_cb_sk_2 Rapid problem-solving & crisis management; swiftly addressing urgent issues to restore stability within a team or organization. Utility lies in mitigating disruptions, maintaining productivity, and preventing escalation of negative situations.
dev_cb_sk_3 Online assessments: Digital tools measuring skills & knowledge through interactive tests and simulations. They provide rapid, scalable insights into candidate abilities, streamlining recruitment and informing talent development strategies.
dev_cb_sk_4 Staying abreast of emerging trends - technologies, methodologies, and evolving consumer preferences - is crucial for adaptability. It enables proactive skill development, career progression, and maintaining a competitive advantage within the dynamic labour market.
dev_cb_sk_5 Preparing tax returns involves accurately calculating and documenting income, deductions, and credits for compliance with tax laws. Its utility lies in minimizing liabilities, maximizing refunds, and ensuring legal adherence to financial reporting standards.
dev_cb_sk_6 Tuning procedures involve adjusting operational processes to optimize efficiency and performance within a workplace. It enhances productivity, reduces waste, and improves overall workflow effectiveness - crucial for competitive labour market success.
dev_cb_sk_7 Iconography methods involve visual representation to communicate complex labour market data - like skills gaps, demographic shifts, or sectoral trends. It's useful for quickly conveying insights and facilitating understanding beyond raw statistics.
```

Figure 2: Skill definitions generated by C_1 ($LLM_1 + Jp_1 + Sp_1$). The skills are taken from the Development Set.

6. Results

The results obtained on the Development Set are presented in Table 1, including those for each individual configuration, the official baseline provided for Task B of TalentCLEF 2025, and the final ensemble. The use of LLMs shows a clear improvement over the official baseline, suggesting that data augmentation—specifically through the inclusion of item definitions—may be beneficial for NLP tasks. Additionally, the improvement observed with the ensemble compared to the individual models indicates that combining multiple systems could help offset the specific limitations of each model.

Table 1
Results obtained on the Development Set. The best results found for each metric are highlighted in bold.

System	MAP	MRR	nDCG	P@5	P@10	P@100
Official Baseline	0.1874	0.6752	0.6660	0.4500	0.3852	0.1964
C_1	0.2162	0.7271	0.6914	0.5053	0.4503	0.2265
C_2	0.2150	0.7039	0.6876	0.4730	0.4390	0.2302
C_3	0.2096	0.7650	0.6888	0.5184	0.4533	0.2195
Ensemble	**0.2442**	**0.7858**	**0.7131**	**0.5539**	**0.4951**	**0.2488**

The Ensemble model, finally, has achieved a MAP of 0.278 in the Test Set, while the official baseline obtained 0.196.

7. Conclusions

The paper *Beyond Titles: Semantic Matching of Jobs and Skills Using LLMs and S-BERT* presents the submission of the AI Research Group in the ITCL Technology Center to TalentCLEF 2025 Task B, where the job-skill matching problem using an ensemble of S-BERT embeddings based on LLM-generated definitions is addressed. The results suggest that data augmentation, through the generation of definitions, can lead to improved performance. Furthermore, combining multiple LLMs may help enhance results, possibly by compensating for the specific limitations of individual models.

The obtained results may be relevant for NLP systems, particularly those that rely on semantic understanding and concept matching, such as the Recommendation Portal currently under development within the AI4Labour project (GA:101007961), on which the ITCL is currently working. Initial evaluations for the development of the algorithms presented in this proposal were conducted using the O*NET database[4], within the scope of that project. This database is the American counterpart to the European ESCO. The Recommendation Portal aims to suggest courses to users based on their skills and educational background, while also seeking to suppress skills that are prone to automation, thereby facilitating adaptation to the AI era. To integrate information from different domains and sources, systems like the one proposed here may be employed.

Acknowledgments

We would like to acknowledge the organizers of TalentCLEF 2025 for providing a well-structured and impactful challenge. Our experiments were conducted using publicly available models and open-source tools. This research was carried out under the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 101007961.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited

[4] https://www.onetonline.org/

the content as needed, and take full responsibility for the publication's content.

References

[1] L. Gasco, H. Fabregat, L. García-Sardiña, D. Deniz, A. Rodrigo, P. Estrella, R. Zbib, Talentclef at clef2025: Skill and job title intelligence for human capital management, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 479–486.

[2] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. S. Mian, A comprehensive overview of large language models, ArXiv abs/2307.06435 (2023). URL: https://api.semanticscholar.org/CorpusID:259847443.

[3] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Conference on Empirical Methods in Natural Language Processing, 2019. URL: https://api.semanticscholar.org/CorpusID:201646309.

[4] 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.

[5] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).