A Union Fusion of PyTerrier, Doc2Query, BM25 and Multilingual Embeddings for Job-Person Matching

TalentCLEF Task A at CLEF 2026

Edwin Thuma[1,*,†], George Anderson[1,†]

[1]Department of Computer Science, University of Botswana, P/BAG UB 00704, Gaborone, Botswana

Abstract

This report details the methodology, pipeline design, and experimental observations for a hybrid retrieval system submitted to the TalentCLEF 2026 Task A competition, by the UBCS (University of Botswana Computer Science) team. The core objective of the task is Contextualized Job-Person Matching, which poses a significant challenge due to the vocabulary mismatch inherent in specialized domains and the added complexity of cross-lingual environments. To address these challenges, we implement a two-pronged approach. First, we establish a robust dense retrieval baseline using a multilingual Siamese transformer embedding model. Second, we employ a sophisticated sparse retrieval pipeline that augments the original documents using an off-line document expansion technique (Doc2Query) prior to indexing and ranking with the BM25 classical probabilistic model. The final ranked lists from both the sparse and dense retrieval strategies are subsequently combined via score-based late fusion, utilizing min-max normalization to handle scale discrepancies. The results of this work suggest that lexical expansion and semantic retrieval provide complementary evidence and that their union can recover a broader and better-ranked set of relevant documents than either component alone.

Keywords

PyTerrier, Doc2Query, BM25, Multilingual Embedding, Union Fusion

1. Introduction

The socio-technological world is changing rapidly due to task automation and Artificial Intelligence (AI). It is creating new roles that demand specialized skills. Language-based systems which use Large Language Models (LLMs) are driving growth of technology in Human Capital Management (HCM) and Human Resources (HR). These new systems are able to process large quantities of data to facilitate the identification of suitable job applicants for available roles based on their resumes [1].

TalentCLEF [2, 3] is a competition that fosters a community approach for a research-oriented approach to solve HCM and HR problems. The specific task addressed in this paper, Task A, is a task in TalentCLEF 2026 [4] seeking to develop models which address the matching of resumes with job openings, addressing issues such as:

(i) Multilingualism: Models are required to understand and process various languages accurately and maintain the context and cultural nuances inherent in each.
(ii) Fairness: Ensuring fairness and reducing bias is a critical challenge in HCM.
(iii) Cross-Industry Adaptability: NLP systems must be flexible enough to work with a variety of sectors such as education, banking, healthcare, and energy.

Task A of the TalentCLEF 2026 competition encourages research in Human Capital Management by establishing rigorous benchmarks for Contextualized Job-Person Matching [5]. The challenge requires participants to identify and rank candidate resumes from a fixed corpus that are semantically equivalent to a given query job offer. This task encompasses both monolingual and cross-lingual settings, testing

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
thumae@ub.ac.bw (E. Thuma); andersong@ub.ac.bw (G. Anderson)
https://www.ub.bw/ (E. Thuma); https://www.ub.bw/ (G. Anderson)
0009-0007-3011-4876 (E. Thuma); 0000-0001-9368-1404 (G. Anderson)

the boundaries of current natural language processing models to capture domain-specific nuances across linguistic boundaries.

Traditional lexical matching approaches can be highly efficient but frequently fail to satisfy information needs when queries and documents utilize disparate vocabularies. This problem is known as vocabulary mismatch; it becomes especially critical in highly specialized settings such as Human Resources and job market platforms. In these contexts, different organizations and geographical locations may employ vastly different terminologies to describe identical occupational roles. For example, the word “doctor” vs the word “physician.”

In this paper, the UBCS (University of Botswana Computer Science) team reports on our approach for solving the stated problem (TalentCLEF Task A at CLEF 2026). To successfully tackle the stated task, we propose a hybrid retrieval architecture that seamlessly integrates the strengths of exact lexical matching with the contextual comprehension of dense semantic embeddings. Our experimental solution utilizes a union fusion pipeline: a baseline multilingual embedding retriever and an advanced sparse retriever augmented by generative document expansion (Doc2Query) [6]. The sparse retriever also served as a baseline. The results from these independent runs are combined through a union-based late fusion process, min-max normalized to ensure balanced influence from both retrieval signals. The performance of this approach improves on our baseline, which makes use of a paraphrase-multilingual-MiniLM-L12-v2 model [7]. We employ the use of Doc2Query [8], BM25 [9], and Embedding Union Retrieval [10]. Therefore, in this research paper, we report on experiments with models developed to find and rank candidates for job positions based on their experience and professional skills.

This research paper is structured as follows. Section 2 provides an overview of the related work concerning document expansion, dense representations, and fusion techniques. Section 3 delineates the experimental methodology, explicitly detailing the operations of the dense retrieval strategy, the sparse retrieval strategy, and the fusion mechanism. Section 4 outlines the execution metrics and simulated outcomes, highlighting the practical aspects of the pipeline such as computational resource tracking. Finally, Section 5 analyzes the advantages and potential drawbacks of the hybrid approach.

2. Related Work

In this section, we summarize relevant literature on sparse retrieval and document expansion, dense retrieval and multilingual embeddings, and information retrieval frameworks and late fusion. Multilingual job title and skill matching has recently become a central focus in NLP for Human Capital Management, driven by new shared tasks, domain-adapted models, and large-scale multilingual embeddings. We also report on some results from the 2025 edition of TalentCLEF.

2.1. Multilingual Job Title Matching Benchmarks and Systems

The TalentCLEF 2025 lab provides the first open benchmark for Multilingual Job Title Matching (Task A), using real, anonymized job applications in English, Spanish, German, and Chinese and explicitly testing monolingual and cross-lingual ranking scenarios as well as gender bias [2]. Most participating systems rely on multilingual encoder-based models fine-tuned with contrastive learning, often combined with large language models (LLMs) for data augmentation or re-ranking, and results indicate that training strategy is more influential than model size alone.

Within this setting, Zhang et al. [11] compare discriminative, contrastive, and prompt-based methods for Task A, showing that large multilingual models with carefully tuned prompts perform best, reaching an average MAP of 0.492 on the TalentCLEF 2025 Task A test set across English, German, and Spanish. The authors compares fine-tuned classification, contrastive, and prompting methods for TalentCLEF. The authors find that large multilingual models and prompt-based setups achieve strong MAP in both mono- and cross-lingual scenarios.

JobBERT-V3 introduces a contrastive-learning encoder trained on over 21 million multilingual titles and reports that it outperforms strong multilingual baselines on TalentCLEF Task A in both monolingual and cross-lingual setups [5]. JobBERT-V3 extends a job-domain encoder to four languages

using synthetic translations and a multilingual job title corpus, outperforming strong multilingual baselines on TalentCLEF and also supporting skill ranking from titles. JobBERT achieved a MAP of 0.6547. These systems illustrate the current state of the art: high-capacity, domain-adapted dense encoders optimized specifically for job title similarity [5].

2.2. Skill Extraction and ESCO-based Alignment

Skill matching work often aligns extracted skills and occupations to the ESCO taxonomy, a multilingual resource with approximately 3,000 occupations and approximately 13,900 skills in 27 languages [12]. ESCOXLM-R uses ESCO-driven domain-adaptive pre-training to build a general multilingual model for job-market tasks such as skill extraction and job title classification, achieving state-of-the-art results on several datasets. Complementarily, a multilingual knowledge-graph framework extracts skills from vacancies and CVs, maps them to ESCO labels, and embeds hierarchical and cross-lingual relations to improve vacancy–candidate matching quality [13].

2.3. Multilingual Embeddings and Domain Adaptation

Several works provide more general language-agnostic sentence embeddings that underpin many of the above systems. Sentence-BERT introduces Siamese BERT networks for efficient semantic similarity search [14]. LaBSE and LASER learn massively multilingual sentence representations across 90+ languages, enabling zero-shot cross-lingual similarity via cosine distance [15, 16]. LaBSE and LASER achieve strong bi-text retrieval and cross-lingual similarity performance with LaBSE reaching 83.7% retrieval accuracy on the Tatoeba benchmark (vs. 65.5% for LASER). Additional methods explicitly distill language-agnostic meaning embeddings from multilingual encoders to improve cross-lingual similarity estimation without task-specific labels [17].

Our findings regarding Sentence-BERT show that Siamese BERT architectures provide efficient semantic similarity search while largely preserving BERT's accuracy, reducing retrieval from approximately 65 hours to seconds on 10,000-sentence collections.

Complementary work constructs a multilingual ESCO-based knowledge graph for vacancy-CV matching and reports significant improvements in matching quality over previous approaches [13].

2.4. Sparse Retrieval and Document Expansion

The Probabilistic Relevance Framework (PRF) forms the backbone of traditional Information Retrieval, culminating in the highly successful BM25 term-weighting model [18]. BM25 excels at fast, exact-match lexical retrieval but inherently struggles with the vocabulary mismatch problem. If a query contains the term "Software Engineer" and the document only uses "Backend Developer," traditional term-matching strategies will fail to establish relevance without extensive, often manually curated, synonym dictionaries.

To circumvent this limitation natively within the sparse retrieval ecosystem, document expansion techniques have become more and more widely used. Instead of enriching the query at run-time, document expansion enriches the document at indexing time. The well-known Doc2Query model [6] revolutionized this domain by training a sequence-to-sequence neural network to predict potential user queries based on a document's content. By appending these generated queries to the original text, the probability of a lexical match during standard BM25 retrieval is significantly increased.

However, generative models are not without flaws. The tendency of sequence-to-sequence models to "hallucinate" irrelevant or incorrect expansions can introduce noise, increasing the index size and occasionally degrading retrieval precision. Recent efforts, such as the Doc2Query– framework [19], attempt to mitigate this by implementing an active filtering mechanism that discards low-quality expansions. However, comprehensive reproducibility studies indicate that while filtering controls index bloat, it can sometimes harm recall-oriented metrics if over-applied [20]. Consequently, our experiment maintains the standard generative expansion approach to maximize coverage, relying on the subsequent fusion stage to suppress noisy outliers.

2.5. Dense Retrieval and Multilingual Embeddings

Dense retrieval models circumvent the vocabulary gap entirely by mapping both queries and documents into a shared continuous vector space, where relevance is computed via simple spatial metrics such as cosine similarity. This capability was primarily unlocked by Sentence-BERT (SBERT), which applied a Siamese network architecture to pre-trained BERT models, allowing for the derivation of semantically meaningful, fixed-size sentence embeddings [14].

In modern setups, embeddings are often further refined through techniques like component focusing to better capture syntactical relationships [21]. For tasks spanning multiple languages, such as TalentCLEF Task A, multilingual variants of these transformers are strictly required. Models like the `paraphrase-multilingual-MiniLM-L12-v2` distill knowledge from larger networks to offer lightweight yet robust representations across dozens of languages, proving highly effective even for short-text topics and low-resource morphologies [22].

2.6. Information Retrieval Frameworks and Late Fusion

Constructing and evaluating complex retrieval pipelines requires flexible software systems. PyTerrier provides a declarative Python-based experimentation framework that allows researchers to efficiently construct pipelines combining classical IR (like BM25) and neural methods [23]. It has been widely utilized in deep learning retrieval tracks to coordinate complex multi-stage setups [24].

In scenarios where multiple heterogeneous ranking signals are available, Information Fusion becomes critical. Late fusion strategies, which combine the final ranked lists of different sub-systems, have long been employed in multimodal environments to merge text, image, and video scores [25, 26]. Our pipeline adopts a score-level linear combination inspired by these foundational fusion models, adjusting for scale differences prior to summation.

2.7. Positioning of Current Work

In contrast to these predominantly dense, contrastively trained systems, the current TalentCLEF 2026 submission adopts a hybrid retrieval strategy: a multilingual Siamese transformer baseline combined with a Doc2Query-expanded BM25 index within PyTerrier, and a simple score-union fusion with min–max normalization. This design aligns with TalentCLEF's retrieval theme but explores a different part of the design space, emphasizing the complementarity of sparse and dense signals rather than further scaling or specializing a single multilingual encoder. The main gap from the above which we have tried to close in our approach is a systematic, reproducible study of Doc2Query-enhanced BM25 plus multilingual dense retrieval with simple, low-cost fusion, tailored to contextualized job-person matching and evaluated on TalentCLEF-style data.

3. Methodology

The objective of this study was to compare lexical expansion, dense semantic retrieval, and a simple fusion strategy under the same evaluation setting. In particular, we deploy the Doc2Query BM25 system, which is a sparse lexical retrieval approach enhanced through document expansion. In addition, we deploy the embedding baseline, which is a dense retrieval strategy based on semantic similarity. Moreover, we deploy the fused union system, which combines the candidate sets produced by the two individual systems in order to exploit complementary retrieval behaviour. The central hypothesis is that the fused union system should improve ranking effectiveness, particularly in terms of MAP, because it has access to relevant documents retrieved by both sparse and dense retrieval mechanisms. The remainder of this section is organized as follows: In Section 3.1, we discuss the preprocessing of data, which is followed by a description of Run 1: Baseline Multilingual Embedding Retrieval in Section 3.2. In Section 3.3 we present Run 2: PyTerrier Doc2Query BM25, which is a sparse lexical retrieval approach. We then employ a union fusion of Run 1 and Run 2 in Section 3.4.

3.1. Dataset and Preprocessing

The experiment is conducted on the TalentCLEF 2026 Task A test set, evaluating job offer matching across specified language pairs, for example English-English and Spanish-Spanish. The provided dataset consists of separate corpus and query directories containing individual files. Each file maps a unique identifier to a specific job offer string.

During initialization, the raw textual data is loaded into structured Pandas DataFrames. A preprocessing pipeline normalizes the text to lowercase and strips extraneous whitespace. This standardizes the lexical footprint for both the dense embedding models and the PyTerrier indexing backend.

3.2. Run 1: Baseline Multilingual Embedding Retrieval

The first run is a dense retriever aiming to maximize semantic matching. We utilize the `paraphrase-multilingual-MiniLM-L12-v2` model loaded via the `sentence-transformers` library. The operation proceeds as follows:

1. **Encoding:** Both the corpus elements and the query strings are passed through the Siamese transformer network to generate 384-dimensional dense vector embeddings.
2. **Similarity Computation:** We compute the exhaustive cosine similarity matrix between the query embeddings and the corpus embeddings.
3. **Ranking:** For each query, the corpus candidates are sorted by their cosine similarity score in descending order, retaining only the top-300 most relevant documents.

This run does not rely on exact keyword matching; instead, it captures latent semantic relationships between related job offers, for example by assigning a high similarity score to "Data Scientist" and "Machine Learning Engineer".

3.3. Run 2: PyTerrier Doc2Query BM25

The second run is designed to provide robust lexical matches, augmented by neural query generation to bridge vocabulary gaps. We employ the PyTerrier framework to orchestrate the workflow:

1. **Document Expansion:** Using a pre-trained sequence-to-sequence model, for example `macavaney/doc2query-t5-base-msmarco`, we generate a fixed number of prospective queries for each job offer in the corpus. These generated terms are appended directly to the original corpus text.
2. **Indexing:** The expanded corpus is indexed using Terrier's inverted indexing architecture. Standard tokenization and stopword removal operations are applied.
3. **Retrieval:** The original (unexpanded) user queries are evaluated against the expanded index using the BM25 weighting function [18]. The system returns a ranked list of top-300 candidates based on the probabilistic scoring.

3.4. Run 3: Union Fusion

Due to the disparate nature of the scores—cosine similarity bounded between $[-1, 1]$ for the dense retrieval run (Run 1), and unbounded probabilistic BM25 scores, we apply score-level late fusion [26]. In particular, the results from Run 1 and Run 2 are first joined using a standard set union. A document retrieved by either system is retained in the final candidate pool. Subsequently, min-max normalization is applied to the scores on a per-query basis:

$$S'_{norm}(d,q) = \frac{S(d,q) - \min_{d \in D_q} S(d,q)}{\max_{d \in D_q} S(d,q) - \min_{d \in D_q} S(d,q)} \tag{1}$$

where $S(d,q)$ is the original score assigned to document d for query q, and D_q is the set of all retrieved documents for query q within that specific run. Documents that were not retrieved by a Run are assigned a normalized score of 0 for that Run.

Table 1
Overall development-set evaluation summary.

Language pair	System	MAP	MRR	Queries evaluated
en-en	fused_union	0.8207	0.9250	10
en-en	doc2query_bm25	0.7586	0.8476	10
en-en	embedding_baseline	0.7059	1.0000	10
es-es	fused_union	0.7585	0.9500	10
es-es	embedding_baseline	0.6470	1.0000	10
es-es	doc2query_bm25	0.6099	0.8083	10

Table 2
Overall test-set evaluation summary.

Language pair	System	MAP	MRR	Queries evaluated
en-en	fused_union	0.513	0.807	40
en-en	embedding_baseline	0.400	0.745	40
es-es	fused_union	0.465	0.782	40
es-es	embedding_baseline	0.364	0.688	40
en-es	fused_union	0.285	0.44	40
en-es	embedding_baseline	0.356	0.714	40

The final fusion score is a weighted linear combination:

$$S_{final} = (\alpha \times S'_{dense}) + ((1 - \alpha) \times S'_{sparse}) \tag{2}$$

For this study, we set equal weights ($\alpha = 0.5$). The final merged result set is sorted, and a ranked list of the top-100 candidate documents is exported in the standard TREC run file format for evaluation.

4. Results

Table 1 reports the overall development-set results for the three systems across the English–English and Spanish–Spanish language pairs. The fused union system achieved the highest MAP in both language pairs. For English–English, the fused union system obtained a MAP of 0.8207, compared with 0.7586 for Doc2Query BM25 and 0.7059 for the embedding baseline. For Spanish–Spanish, the fused union system obtained a MAP of 0.7585, compared with 0.6099 for Doc2Query BM25 and 0.6470 for the embedding baseline.

Table 2 presents the overall test-set evaluation results. In the monolingual settings, the `fused_union` system consistently outperformed the `embedding_baseline`. For English–English, `fused_union` achieved a MAP of 0.513 and an MRR of 0.807, compared with 0.400 and 0.745 for the embedding baseline. A similar trend was observed for Spanish–Spanish, where `fused_union` obtained a MAP of 0.465 and an MRR of 0.782, while the embedding baseline achieved 0.364 and 0.688, respectively. These results indicate that combining lexical and semantic retrieval signals improves ranking effectiveness in monolingual job title matching.

However, the cross-lingual English–Spanish results show a different pattern. In this setting, the embedding baseline achieved a higher MAP of 0.356 and an MRR of 0.714, compared with 0.285 and 0.440 for the fused union system. This suggests that the dense multilingual embedding model was better able to capture cross-lingual semantic similarity, whereas the lexical component of the fused system may have introduced weaker or less transferable matches across languages. Overall, the test-set results support the usefulness of fusion for monolingual retrieval, but they also show that cross-lingual matching requires stronger reliance on multilingual semantic representations.

5. Discussion

The test-set and the development-set results show that the fused union system is the strongest configuration when effectiveness is measured using MAP. This pattern is consistent across both language pairs (en-en and es-es), suggesting that combining Doc2Query BM25 with embedding-based retrieval improves the overall ranking of relevant documents. These results support the hypothesis that lexical expansion and semantic retrieval provide complementary evidence, and that their union can recover a broader and better-ranked set of relevant documents than either component alone. However, the English–Spanish cross-lingual test-set results reveal a different trend. In this setting, the embedding baseline achieved a higher MAP of 0.356 and an MRR of 0.714, compared with 0.285 and 0.440 for the fused system. This indicates that the multilingual embedding model is more effective than the fused approach when the query and candidate job offers are in different languages. The weaker performance of the fused system in this setting suggests that the lexical BM25-based component contributes less useful evidence across languages and may introduce noisy candidates into the union. Since BM25 depends heavily on token overlap, its benefits are reduced when equivalent job offers are expressed in different languages.

6. Conclusion

This work set out to investigate whether the fused union system could improve the ranking effectiveness, particularly in terms of MAP, by combining relevant documents retrieved by both sparse and dense retrieval mechanisms. Overall, the test-set results confirm that fusion is effective for monolingual retrieval, where lexical and semantic signals can reinforce each other. At the same time, they show that cross-lingual job-person matching requires stronger reliance on multilingual semantic representations. Future work should therefore consider language-aware fusion, where the lexical component is weighted more heavily for monolingual pairs but reduced or replaced in cross-lingual settings. Another possible improvement would be to include translation-based expansion, multilingual Doc2Query, or a cross-lingual reranking stage to improve the quality of fused results for English–Spanish matching.

Acknowledgments

The authors thank the organizers of TalentCLEF 2026 for a well-organized research competition.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.5 and Gemini 3.1 for brain-storming and English language editing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] TalentCLEF organisers, TalentCLEF: About TalentCLEF, https://talentclef.github.io/talentclef/about/, 2026. Accessed: 2026-05-19.

[2] 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, 2025, pp. 464–485.

[3] 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.

[4] 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.
[5] J.-J. Decorte, M. De Lange, J. Van Hautte, Multilingual jobbert for cross-lingual job title matching, arXiv (2025). doi:10.48550/arxiv.2507.21609.
[6] R. Nogueira, W. Yang, J. Lin, K. Cho, Document expansion by query prediction, arXiv (2019). doi:10.48550/arxiv.1904.08375.
[7] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021). doi:10.18653/v1/2021.findings-acl.188.
[8] R. Nogueira, J. Lin, From doc2query to docTTTTTquery, 2019. URL: https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery.pdf, _eprint: 1904.08375.
[9] S. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.
[10] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2009, pp. 758–759. URL: https://doi.org/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.
[11] M. Zhang, R. van der Goot, NLPnorth @ TalentCLEF 2025: Comparing discriminative, contrastive, and prompt-based methods for job title and skill matching, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025, pp. 4525–4535. URL: https://ceur-ws.org/Vol-4038/paper_377.pdf.
[12] M. Zhang, R. van der Goot, B. Plank, Escoxlm-r: Multilingual taxonomy-driven pre-training for the job market domain, ArXiv abs/2305.12092 (2023). doi:10.48550/arxiv.2305.12092.
[13] H. Kavas, M. Serra-Vidal, L. Wanner, Multilingual skill extraction for job vacancy-job seeker matching in knowledge graphs (2025) 146–155.
[14] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. doi:10.18653/v1/d19-1410.
[15] F. Feng, Y. Yang, D. M. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding (2020) 878–891. doi:10.18653/v1/2022.acl-long.62.
[16] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2018) 597–610. doi:10.1162/tacl_a_00288.
[17] N. Tiyajamorn, T. Kajiwara, Y. Arase, M. Onizuka, Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation (2021) 7764–7774. doi:10.18653/v1/2021.emnlp-main.612.
[18] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.
[19] M. Gospodinov, S. MacAvaney, C. Macdonald, Doc2query–: When less is more, in: Lecture Notes in Computer Science, Springer Nature Switzerland, 2023, pp. 414–422. doi:10.1007/978-3-031-28238-6_31.
[20] W. Mansour, S. Zhuang, G. Zuccon, J. Mackenzie, Revisiting document expansion and filtering for effective first-stage retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2024, pp. 186–196. doi:10.1145/3626772.3657850.
[21] X. Yin, W. Zhang, W. Zhu, S. Liu, T. Yao, Improving sentence representations via component focusing, Applied Sciences 10 (2020) 958. doi:10.3390/app10030958.
[22] D. Medvecki, B. Bašaragin, A. Ljajić, N. Milošević, Multilingual transformer and bertopic for short text topic modeling: The case of serbian, in: Disruptive Information Technologies for a Smart Society. ICIST 2023, Springer, Cham, 2024. doi:10.1007/978-3-031-50755-7_16.

[23] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, 2020. doi:10.1145/3409256.3409829.

[24] X. Wang, Y. Wu, X. Wang, C. Macdonald, I. Ounis, University of Glasgow Terrier Team at the TREC 2020 Deep Learning Track, Technical Report, National Institute of Standards and Technology (NIST), 2020. doi:10.6028/nist.sp.1266.deep-uogtr.

[25] A. Depeursinge, H. Müller, Fusion techniques for combining textual and visual information retrieval, in: The Information Retrieval Series, Springer Berlin Heidelberg, 2010, pp. 95–114. doi:10.1007/978-3-642-15181-1_6.

[26] C. Moulin, C. Largeron, C. Ducottet, M. Géry, C. Barat, Fisher linear discriminant analysis for text-image combination in multimedia information retrieval, Pattern Recognition 47 (2014) 260–269. doi:10.1016/j.patcog.2013.06.003.