Unite and Conquer: WebNLG-IR for Evaluating Triple, Text, and Hybrid Retrieval over Multilingual Knowledge Graphs

Virginia Ramón-Ferrer[1,*], Carlos Badenes-Olmedo[1] and Oscar Corcho[1]

[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

Abstract

Knowledge graph (KG) retrieval is often addressed through formal query languages such as SPARQL, but such access may be impractical when users lack schema knowledge or maintained endpoints are unavailable. As an alternative, we study whether KG content can be retrieved through natural-language representations. Existing multilingual IR benchmarks generally do not provide aligned triple and verbalised representations of the same KG facts for controlled retrieval evaluation. We therefore introduce WebNLG-IR, an IR benchmark built from the established WebNLG dataset, with English, Spanish, and Catalan queries and English and Spanish documents represented as triple sets, lexicalisations, or their combination. We evaluate several retrieval paradigms across different complexity levels, defined by the number of triples in each entry. Our results show that combining triple and text evidence improves retrieval, with simple textual concatenation achieving the strongest overall performance. This suggests that data-to-text verbalisation provides useful support for bridging structured KG facts and multilingual natural-language queries.

Keywords

Knowledge Graph Retrieval, Multilingual Information Retrieval, Data-to-Text Verbalisation, WebNLG

1. Introduction

Knowledge graphs (KGs) provide structured representations of factual knowledge and are commonly accessed through formal query languages such as SPARQL [1]. However, this access model assumes that users can formulate formal queries and understand the structure, entities, and predicates of the underlying graph, which makes direct KG access difficult for non-expert users [2, 3]. Moreover, deploying and maintaining a SPARQL-based access layer may introduce additional modelling, engineering, and maintenance costs, which may not always be justified for smaller-scale retrieval scenarios or for KGs without a stable formal schema. These limitations have motivated alternative access paradigms based on natural-language questions, keyword search, and information retrieval (IR) over KG-derived content. Prior work has explored IR-oriented access to RDF and KG data by indexing triples, entities, subgraphs, or virtual documents as textual retrieval units [4, 5]. More recent KGQA and GraphRAG approaches also retrieve KG facts after converting them into textual forms, either as simple subject–predicate–object strings or as natural-language verbalisations [6, 7]. In parallel, GraphRAG and hybrid-retrieval work has highlighted the importance of combining relational and textual evidence for retrieval-augmented systems [8, 9, 10]. However, they do not directly isolate retrieval effects while keeping the underlying KG facts fixed and varying their representation as triples, verbalised text, or both, particularly in multilingual settings.

In this paper, we address this gap by introducing WebNLG-IR, a multilingual IR benchmark derived from WebNLG and its Spanish adaptation [11, 12]. It contains English, Spanish, and Catalan queries, and

The 25th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2026) - Satellite Events (Workshops, Tutorials, Demos, Posters).

*Corresponding author.

virginia.ramon@upm.es (V. Ramón-Ferrer); carlos.badenes@upm.es (C. Badenes-Olmedo); oscar.corcho@upm.es (O. Corcho)

0009-0009-7197-1676 (V. Ramón-Ferrer); 0000-0002-2753-9917 (C. Badenes-Olmedo); 0000-0002-9260-0753 (O. Corcho)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-26

CEUR-WS.org/Vol-4340/paper4.pdf

English, Spanish, and mixed-language documents represented as triples, verbalised text, or combinations of both. This enables direct comparison of the same KG facts represented as textual triples, natural-language verbalisations, concatenated triple–text documents, or hybrid representations obtained by combining separate triple and text retrieval results. Our experiments evaluate these representations across several retrieval paradigms and document-complexity levels, defined by the number of triples in each document. The results show that triples generally outperform verbalised text alone, suggesting that their compact relational structure remains useful for retrieval. However, the best results are obtained by concatenating triples and verbalised text into a single representation. This indicates that verbalisation is useful when combined with structured triples, as it provides additional natural-language cues while preserving the explicit facts encoded in the KG.

2. Related Work

Textual information retrieval over knowledge graphs has been studied as an alternative to formal query-based access, especially when users cannot be expected to write SPARQL queries or know the schema of the underlying graph. Prior work on RDF and KG search has represented structured data as retrievable textual units, including triples, entities, subgraphs, and virtual documents, enabling the use of standard IR models over graph-derived content [4, 5]. These approaches connect KG retrieval with broader textual IR methods, ranging from lexical sparse models such as BM25, which rely on exact term matching and document-length normalisation [13], to dense bi-encoder models such as BGE, which encode queries and documents as vectors for semantic matching [14]. Recent multilingual sparse models such as MILCO combine sparse retrieval with cross-lingual matching, while multivector models represent each query and document with multiple vectors, allowing finer-grained token-level or late-interaction matching [15, 14]. However, while these retrieval paradigms provide different ways of matching natural-language queries to documents, they do not directly isolate retrieval effects while holding the underlying KG facts fixed and varying their representation as triples, verbalised text, or both.

Data-to-text generation provides a complementary line of work in which structured triples are converted into natural-language text. The WebNLG challenge established a widely used benchmark for generating textual descriptions from RDF triples [11], and later adaptations extended this setting to other languages [16, 12]. WebNLG has also been used to derive QA resources, including WEBNLG-QA/WebNLQ-QA for conversational KGQA with SPARQL queries [17], and multilingual graph–text QG/QA data for generating and answering questions from KGs and texts [18]. Existing benchmarks include multilingual IR resources such as MIRACL [19] and WebFAQ [20], and semi-structured retrieval benchmarks such as STaRK [10]. However, they do not provide aligned triple and verbalised representations of the same KG facts and therefore cannot directly support this controlled comparison. Similarly, WebNLG-derived QA/QG resources target question answering or generation rather than ranking KG-derived documents under alternative representations. More recent KGQA and GraphRAG approaches retrieve KG facts as triple-form strings, verbalised statements, or hybrid textual–relational evidence for downstream QA and generation [6, 7, 8, 9, 10]. WebNLG-IR fills this gap by providing a multilingual IR benchmark that uses single-triple questions as controlled information needs, projects them onto larger multi-triple documents, and systematically compares triples, verbalised text, concatenated triple–text documents, and late-fusion hybrids across languages, retrieval methods, and document-complexity levels.

3. WebNLG-IR Dataset

We build WebNLG-IR[1] from WebNLG and its Spanish adaptation [11, 12]. We generate extractive and yes/no question–answer pairs from single-triple lexicalisations using prompted LLMs, so each

[1]Dataset and code are available in https://github.com/virginia-r99/WebNLG-IR.

query is linked to one source triple. A validation step regenerates invalid pairs. The validated English questions are translated into Spanish and Catalan using the corresponding lexicalisations as grounding context. Catalan provides an additional query language not represented among the document languages. A document is relevant when its tripleset contains the source triple used to generate the query, allowing single-triple queries to evaluate multi-triple documents with consistent relevance across representations. Documents are represented as textual triples, WebNLG lexicalisations, or mixed-language triples. The mixed-language condition randomly selects the English or Spanish version at each aligned triple position while preserving the tripleset structure.

Since WebNLG-IR is constructed automatically, human evaluation of the generated questions, answers, and relevance judgments remains for future work. However, the benchmark is intended primarily for comparing document representation methods under the same experimental conditions, rather than as an absolute measure of retrieval performance, giving all representations a consistent evaluation setting. Although answer strings are included to trace the intended fact and support future QA extensions, they are not exhaustively validated, as it is not the primary goal of this work. Table 1 summarises the dataset: 35,616 prepared queries, 16,631 canonical documents, and 503,985 relevance judgments across English, Spanish, and Catalan queries, with English, Spanish, and mixed-language document conditions. It also reports 17,451 extractive, 9,093 negative yes/no, and 9,072 positive yes/no queries; the slight imbalance is due to the automatic validation phase, which filtered out queries regarded as invalid or unreliable.

Table 1
Summary of the dataset.

Statistic	Value
Prepared queries	35,616
Canonical documents	16,631
Qrel entries	503,985

Doc size	# Docs
≥ 2	12,664
≥ 3	9,530
≥ 4	6,125

Query type	# Queries
Extractive	17,451
No	9,093
Yes	9,072

Query languages: en, ca, es. Document languages / conditions: en, es, mix.

4. Experimental Setup

We evaluate representation performance on WebNLG-IR across query–document language combinations and document-complexity levels, as summarised in Table 1. Queries are issued in English, Spanish, and Catalan, and paired with English, Spanish, and mixed-language document conditions, allowing us to test whether representation effects are stable across monolingual, cross-lingual, and mixed-language[2] settings. To analyse document complexity, we evaluate subsets containing documents with at least 2, 3, and 4 triples. These thresholds cover a controlled complexity range while avoiding higher thresholds, where the document pool becomes too small for reliable comparison. For each level, we compare five document representations: triples alone, verbalised text alone, concatenated textual triples and text, hybrid reciprocal rank fusion (RRF) over separate triple and text rankings [21], and hybrid score-level fusion over separate triple and text scores. Unlike concatenation, the hybrid methods retrieve triples and text separately. RRF combines their rankings with equal weights and $k = 60$, while score fusion averages per-query min–max normalised triple and text scores.

We evaluate retrieval effectiveness using nDCG@10 (normalised discounted cumulative gain at rank 10) [22], a standard IR metric that rewards highly ranked relevant documents through a logarithmic rank discount and normalises scores by the ideal ranking for each query. This supports comparison across queries with different numbers of relevant documents. We also test four retrieval methods:

[2]Mixed-language denotes triple sets containing both English and Spanish triples. We use English as the corresponding text for hybrid representations.

BM25, a lexical sparse model based on exact term matching and document-length normalisation [13]; BGE dense retrieval, which ranks single-vector query and document embeddings by similarity [14]; MILCO, a learned multilingual sparse retriever [15]; and BGE multivector retrieval, which uses multiple vectors for finer-grained late-interaction matching [14]. Our goal is not to rank retrieval models, but to use them as methods for evaluating and contrasting document representations. We analyse whether lexical, dense, learned sparse, and multivector methods show consistent preferences for triples, text, concatenated, or hybrid representations, and whether these preferences vary by retrieval method, language condition, or document complexity. The code and results obtained from this evaluation are available in https://github.com/virginia-r99/WebNLG-IR.

5. Results



(a) ≥2 (b) ≥3 (c) ≥4

Figure 1: Retrieval method results per document representation and complexity level

Figure 1 shows the nDCG@10 results across retrieval methods and document-complexity settings. The same pattern holds for ≥ 2, ≥ 3, and ≥ 4 triples: verbalised text alone is generally the weakest representation, triples consistently outperform text, and concatenating textual triples with verbalised text achieves the best results. This trend is observed for all retrieval methods, including BM25, BGE dense retrieval, BGE multivector retrieval, and MILCO. For instance, MILCO reaches around 0.82–0.83 nDCG@10 with concatenation, compared with around 0.76–0.77 for triples and 0.72–0.73 for text. Although BM25 obtains lower absolute scores, concatenation remains its strongest representation. The hybrid methods also improve over text alone, but generally do not surpass concatenation. This suggests that retrieving triples and text separately and combining their rankings or scores is less effective than exposing both sources of evidence jointly to the retriever. Score fusion is usually slightly better than RRF, but both remain close to the triples-only baseline and below concatenation. Thus, the main benefit appears to come from unifying structured and verbalised evidence within a single document representation rather than from late fusion alone. Figure 2 shows that this behaviour is consistent across query–document language combinations. Concatenation achieves the highest nDCG@10 in all language directions, including monolingual, cross-lingual, and mixed-language settings. This indicates that the advantage of combining triples with verbalised text is not restricted to a specific language pair, but is consistent across the multilingual conditions considered in WebNLG-IR. The same pattern is also observed across question types. For extractive, positive yes/no, and negative yes/no queries, concatenation remains the strongest representation, indicating that the benefit of combining triples and verbalised text is not specific to a particular query form.

We hypothesise that triples outperform verbalised text because they provide a compact and explicit representation of each fact. By directly encoding the subject, relation, and object, triples reduce linguistic variation and peripheral wording, making relevant entities and relations easier to match. In contrast, verbalised text may introduce paraphrases, discourse structure, or lexical choices that make the supporting relation less explicit, which may explain why text alone performs worse despite being more natural. We further theorise that concatenation works best because triples and verbalised text provide complementary signals. Triples expose the structured fact in a concise form, while verbalised

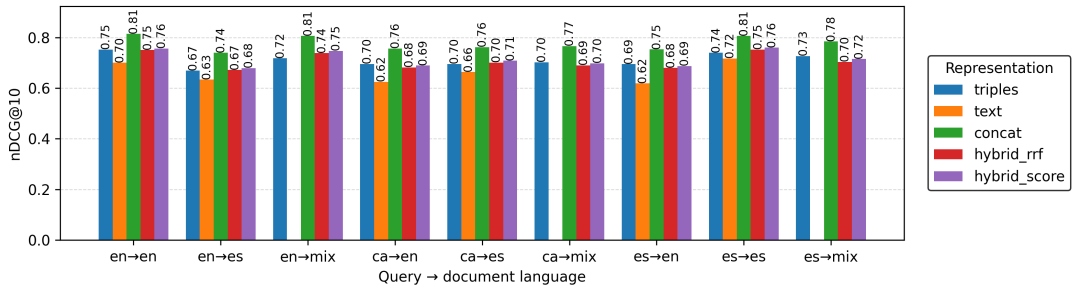


Figure 2: Document representation by language direction for document complexity ≥2. English is used as the corresponding text for the mixed-language condition.

text adds natural-language context and alternative lexical cues that may better match some queries. Concatenation allows retrieval models to use both signals jointly within the same document, whereas score fusion can only combine evidence after separate retrieval runs. This may explain why score fusion improves over triples in some cases but is less consistently effective than concatenation.

6. Conclusions and Future Work

We presented WebNLG-IR, a multilingual information retrieval benchmark derived from WebNLG and its Spanish adaptation, designed to evaluate how different document representations support retrieval over structured knowledge. The dataset contains English, Spanish, and Catalan queries, English, Spanish, and mixed-language document conditions, and relevance judgments projected from single-triple questions to more complex documents. Using this benchmark, we compared triples, verbalised text, concatenated triples and text, and two hybrid fusion strategies across several retrieval paradigms and document-complexity levels. Within this controlled setting, the results show that verbalisation is most useful when combined with triples: verbalised text alone underperforms, while concatenating it with triples yields the strongest performance. These results suggest that triples and verbalised text contribute complementary signals. Triples preserve compact relational structure and may help avoid noise introduced by natural-language paraphrasing, while verbalised text makes the same facts more accessible to lexical, dense, and multilingual retrieval models. The strong performance of concatenation suggests that exposing both forms in a single retrievable document may be more effective than combining separate rankings after retrieval, since the model can match against both structured and natural-language evidence simultaneously. Future work aims to extend WebNLG-IR with human validation of questions, answers, and relevance judgments, additional languages and knowledge graph domains, and more detailed analyses of query type, mixed-language robustness, and retrieval behaviour over larger and more complex graph-derived documents.

Acknowledgments

This work has been supported by the IARAG (CPP2023-010895) project, funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style, and Content enhancement, as well as to suggest refinements to the authors' code. After using this tool, the authors reviewed and edited the content and code as needed and take full responsibility for the publication's content.

References

[1] S. Harris, A. Seaborne, SPARQL 1.1 query language, W3C Recommendation, 2013. URL: https://www.w3.org/TR/sparql11-query/.

[2] H. M. Zahera, M. Ali, M. A. Sherif, D. Moussallem, A.-C. N. Ngomo, Generating sparql from natural language using chain-of-thoughts prompting., in: SEMANTICS, 2024, pp. 353–368.

[3] J. D'Abramo, A. Zugarini, P. Torroni, Investigating large language models for text-to-SPARQL generation, in: W. Shi, W. Yu, A. Asai, M. Jiang, G. Durrett, H. Hajishirzi, L. Zettlemoyer (Eds.), Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing, Association for Computational Linguistics, Albuquerque, New Mexico, USA, 2025, pp. 66–80. URL: https://aclanthology.org/2025.knowledgenlp-1.5/. doi:10.18653/v1/2025.knowledgenlp-1.5.

[4] G. Kadilierakis, P. Fafalios, P. Papadakos, Y. Tzitzikas, Keyword search over rdf using document-centric information retrieval systems, in: The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2020, p. 121–137. URL: https://doi.org/10.1007/978-3-030-49461-2_8. doi:10.1007/978-3-030-49461-2_8.

[5] D. Dosso, G. Silvello, A scalable virtual document-based keyword search system for rdf datasets, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 965–968. URL: https://doi.org/10.1145/3331184.3331284. doi:10.1145/3331184.3331284.

[6] J. Baek, A. F. Aji, A. Saffari, Knowledge-augmented language model prompting for zero-shot knowledge graph question answering, in: B. Dalvi Mishra, G. Durrett, P. Jansen, D. Neves Ribeiro, J. Wei (Eds.), Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 78–106. URL: https://aclanthology.org/2023.nlrse-1.7/. doi:10.18653/v1/2023.nlrse-1.7.

[7] Y. Wu, N. Hu, S. Bi, G. Qi, J. Ren, A. Xie, W. Song, Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering, 2023. URL: https://arxiv.org/abs/2309.11206. arXiv:2309.11206.

[8] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, 2024. URL: https://arxiv.org/abs/2408.08921. arXiv:2408.08921.

[9] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, S. Pasquali, Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction, in: Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 608–616. URL: https://doi.org/10.1145/3677052.3698671. doi:10.1145/3677052.3698671.

[10] S. Wu, S. Zhao, M. Yasunaga, K. Huang, K. Cao, Q. Huang, V. N. Ioannidis, K. Subbian, J. Zou, J. Leskovec, Stark: Benchmarking llm retrieval on textual and relational knowledge bases, 2024. URL: https://arxiv.org/abs/2404.13207. arXiv:2404.13207.

[11] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, The webnlg challenge: Generating text from rdf data, in: 10th International Conference on Natural Language Generation, ACL Anthology, 2017, pp. 124–133.

[12] V. Ramón-Ferrer, C. Badenes-Olmedo, O. Corcho, Spanish triple-to-text benchmark on low-resource large language models, Semantic Web 17 (2026) 22104968261440701.

[13] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[15] T. Nguyen, Y. Lei, J.-H. Ju, E. Yang, A. Yates, Milco: Learned sparse retrieval across languages via a multilingual connector, in: International Conference on Learning Representations, 2026.

[16] M. Oliverio, P. F. Balestrucci, A. Mazzei, V. Basile, Webnlg-it: Construction of an aligned rdf-italian corpus through machine translation techniques, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 12073–12083.

[17] G. Lecorvé, M. Veyret, Q. Brabant, L. M. Rojas Barahona, SPARQL-to-text question generation for knowledge-based conversational applications, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 131–147. URL: https://aclanthology.org/2022.aacl-main.11/. doi:10.18653/v1/2022.aacl-main.11.

[18] K. Han, C. Gardent, Multilingual generation and answering of questions from texts and knowledge graphs, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 13740–13756. URL: https://aclanthology.org/2023.findings-emnlp.918/. doi:10.18653/v1/2023.findings-emnlp.918.

[19] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, J. Lin, Making a miracl: Multilingual information retrieval across a continuum of languages, 2022. URL: https://arxiv.org/abs/2210.09984. arXiv:2210.09984.

[20] M. Dinzinger, L. Caspari, K. Ghosh Dastidar, J. Mitrović, M. Granitzer, Webfaq: A multilingual collection of natural qa datasets for dense retrieval, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 3802–3811. URL: https://doi.org/10.1145/3726302.3731934. doi:10.1145/3726302.3731934.

[21] G. V. Cormack, C. L. A. Clarke, S. Buettcher, 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, SIGIR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 758–759. URL: https://doi.org/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.

[22] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: https://doi.org/10.1145/582415.582418. doi:10.1145/582415.582418.