ARSHAD, U., CORSAR, D. and NKISI-ORJI, I. 2024. Integrating KGs and ontologies with RAG for personalised summarisation in regulatory compliance. In Martin, K., Salimi, P. and Wijayasekara, V. (eds.) 2024. *SICSA REALLM workshop 2024: proceedings of the SICSA (Scottish Informatics and Computer Science Alliance) REALLM (Reasoning, explanation and applications of large language models) workshop (SICSA REALLM workshop 2024), 17 October 2024, Aberdeen, UK*. CEUR workshop proceedings, 3822. Aachen: CEUR-WS [online], pages 56-61. Available from: https://ceur-ws.org/Vol-3822/short7.pdf

Integrating KGs and ontologies with RAG for personalised summarisation in regulatory compliance.

ARSHAD, U., CORSAR, D. and NKISI-ORJI, I.

2024

Integrating KGs and Ontologies with RAG for Personalised Summarisation in Regulatory Compliance

Umair Arshad[1,*,†], David Corsar[2] and Ikechukwu Nkisi-Orji[3]

[1,2,3]School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen AB10 7GJ, Scotland, United Kingdom

Abstract

With the growing complexity and increased volumes, regulatory texts are fast becoming a significant challenge for organisations to remain compliant. Traditional ways of summarising legal texts need to be more accommodating of critical, domain-specific requirements, rendering the process ultimately inefficient and subject to the risk of non-compliance. Therefore, this paper proposes a new solution integrating Ontology and Knowledge Graphs (KGs) with the Retrieval-Augmented Generation (RAG) paradigm to aid process automation and improve regulatory compliance. It offers deep semantic understanding, accurate contextual summaries, and personalised insights relevant to users' needs. In the meantime, this will assist organisations in operating with more precision and confidence in an ever-changing regulatory environment.

Keywords

Regulatory Compliance, KGs, RAG, Personalised Summarisation

1. Introduction

Regulations are becoming more complex and numerous, and with a greater frequency of change, compliance is becoming a continuing challenge for companies. In line with this, compliance teams must constantly adapt to meet evolving regulatory requirements [1]. They are essentially managing regulatory documents, finding the right content, and matching that to their organisation's situation, which is a complex, manual, and inefficient process [2]. The process could benefit greatly from enhanced capabilities in natural language understanding technologies through automated provision of support based on identifying and summarising parts of regulatory documents related to the specific task at hand. Most research into text summarisation of legal documents has been on extractive approaches, which assemble vital phrases from a text to produce a summary [3]. However, these techniques are ineffective in domains with closed vocabulary, including legal regulatory documents, where capturing semantic relationships and contextual dependencies is essential [4]. However, these documents will likely have complex clauses, conditions, and implicit references to be interpreted in-depth, more than the surface level of extraction needed.

Therefore, more advanced natural language processing (NLP) techniques will be needed to account for the subtleties of the regulatory documents, considering the implicit semantic relations expressed throughout both in terms of understanding the documents and generation of outputs, e.g., summaries. Ontologies, as a well-established methodology, describe the concepts within a domain and the relationships between them. Ontologies provide a schema for knowledge graphs (KGs), which capture individual examples of the concepts and their interconnections. The domain knowledge captured in ontologies and KGs can be successfully used to support question-answering tasks [5]. While creating ontologies and knowledge graphs is a complex process that demands significant domain knowledge and can be particularly challenging in changing environments, such as regulations, approaches to automated KG construction for extracting knowledge and structure from unstructured documents have potential to reduce the ontology maintenance overheads [6, 7, 8].

Combining recent advances in NLP with ontologies specific to the regulatory domain, there is potential to improve the effectiveness of systems that support compliance teams within organisations. One such advance, is Retrieval-Augmented Generation (RAG) architectures. RAG combine representations

SICSA REALLM Workshop 2024

U.Arshad1@rgu.ac.uk (U. Arshad); d.corsar1@rgu.ac.uk (D. Corsar); i.nkisi-orji@rgu.ac.uk (I. Nkisi-Orji)

of unstructured information in the form of text embeddings, with additional knowledge – such as from ontologies – to improve outputs when compared to those of Large Language Models (LLMs) on similar tasks [9]. Learning how to generate embeddings, of both the data (in this case, regulatory documents) and user's query (e.g. a request for summary of updates) is a critical part of this process. This enables the system to retrieve more accurate results when searching for potential answers, by utilising the domain knowledge in the ontology. To support this process, [10] demonstrated how the use of a cross-attention mechanism can be used to allows the system to weigh in and prioritise the retrieved information, according to relevance to the query to highlighting the most important details in the outputs. This approach aims to address some of the challenges that LLMs face when analysing domain-specific content [11]. The KG can also be used to evaluate the system's output, providing domain constraints that can be used to assess the factual accuracy of the generated outputs.

2. Proposed Approach

This position paper proposes combining regulatory domain knowledge expressed in ontologies and KGs, with the text understanding and generation capabilities of LLMs within a RAG architecture to enhance the automated support provided with monitoring regulatory compliance. Figure 1 outlines the proposed architecture. The first step in the system is data preprocessing: this step will take inputs from different sources, such as ontologies and regulation documentation files, and perform any necessary the necessary data cleaning and processing steps to create and store embeddings of the domain knowledge in each source. Next, when a user's query is received, a query embedding is obtained by mapping the it using a pre-trained language model, such as BERT [12]. These process the query into syntactic and semantic parts, making a vector representation of the query - necessary for retrieving relevant information from the data sources. The query embedding is then used to search both in a Vector Database of the unstructured text and of the KG (i.e. structured data). Vector databases, such as Milvus [13], are used to quickly retrieve relevant regulatory text; a Neo4j graph database is used to store the ontology and knowledge graph, supporting retrieval of relevant concepts and entities [14]. Graph embedding techniques like TransE can be used to generate embeddings for entities and relationships in KG and transform them into vectors [15]. Then, cosine similarity or other similarity metrics are used to compare the user query embedding with KG and text embeddings to retrieve information that is contextually relevant. This dual retrieval process is used to ensure the returned information is appropriate and relevant to the user's query, through the combination of through structured and unstructured information about regulations.

After the retrieval of relevant information, the system performs a retrieval evaluation to rank according to the relevance and importance of data retrieved. For each retrieved piece of information, a cross attention mechanism [11] can be used to calculate an attention score providing an indication of how much the retrieved information correlates with the query defined by the user. A context vector consisting of the most relevant features from the knowledge graph and retrieved text is generated by pooling the highest-ranked results. Then, this context vector is processed with Transformer Layers that refine the information further, and generate a response aligned with the query.

To improve performance, the generated response has to go through an evaluation step before being presented to the user. In the proposed approach, this step involves checking the accuracy of the information within the response using domain-specific rules and legal standards stored in the KG. These are constraints to make sure that the output is valid is a semantically accurate presentation of the legal information in the regulations, and is appropriate for the context. This can be achieved using methodologies such as those suggested by [16, 17]. If the response is validated, it gets enriched with structured insights from the KG as a personalised summary and facilitates the final stage. The system iterates and revises the response if the validation criteria are not met, safeguarded by iteration limits to prevent infinite loops. It assures that only high-quality, legally compliant information is passed to the user.

To assess the performance of the proposed solution, several evaluation metrics will be employed:

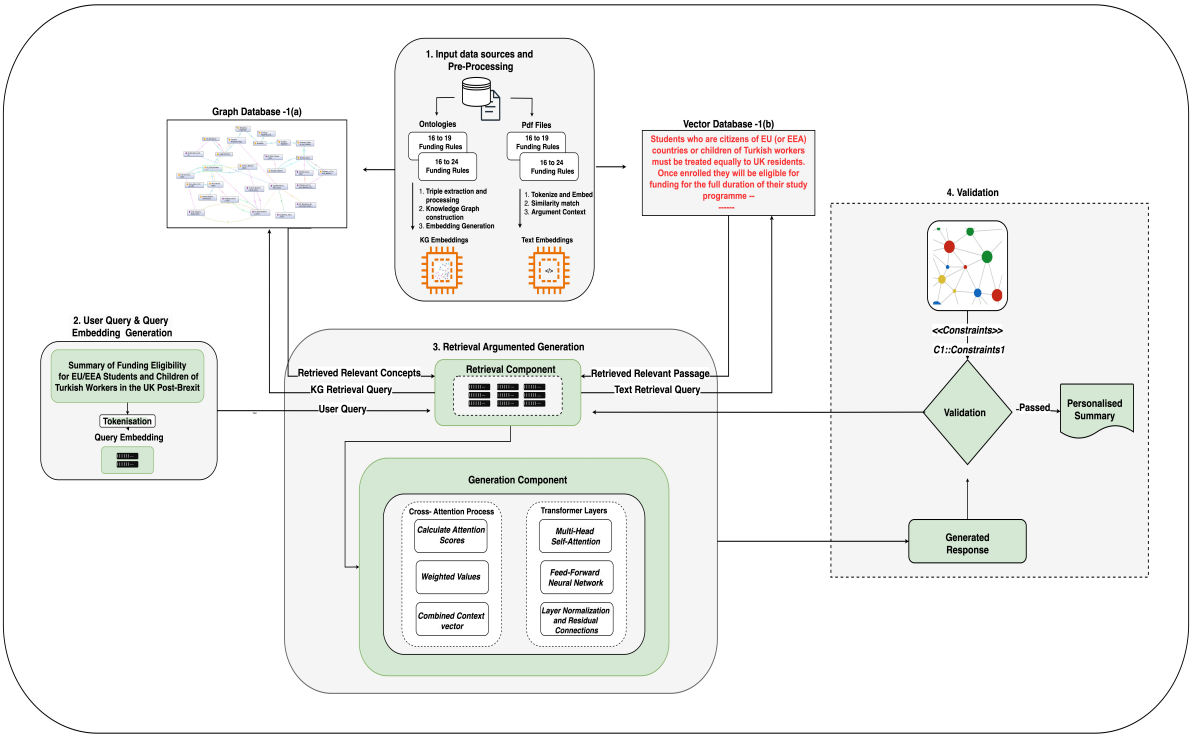

Figure 1: Architecture of Enhanced Personalised Summarisation of Regulatory Compliance

- **Accuracy** will be measured by checking the factual correctness of the output against the corresponding regulations.
- **Relevance** will be assessed using attention scores from the cross-attention mechanism to ensure the retrieved information aligns contextually with the query.
- **Completeness** will be evaluated through manual reviews by domain experts, ensuring that the summary addresses all critical aspects of the query.
- **User satisfaction** will be gauged through surveys and feedback from end-users (e.g., compliance officers) using metrics like Net Promoter Score (NPS).
- **Efficiency** will be tested by benchmarking the system's response time and scalability performance.

By integrating ontologies, KG, and LLMs in a RAG architecture, and evaluating performance using the outlined metrics, this solution promises to simplify the process of monitoring regulatory compliance.

3. Rationale for the Proposed Solution

The proposed approach will be far more effective in managing complications arising from regulatory compliance by automating conventional time- and labour-intensive processes and enhancing scalability and efficiency [1]. Such automation is required, given that the regulatory landscape keeps expanding, and it is hard for the manual effort to keep pace. Instead, what will be had is an application of ontologies and integration of KGs—the model that has a deeper grasp of the legal concepts and how they interlink. In that respect, it will offer the generated summaries relevantly specific in content, retaining better representation in the regulatory documents [18]. By so doing, it will serve the specific legal professionals and organisations better and make the whole process reliable and more comprehensive.

It is also more agile since RAG will remain responsive to an ever-changing rules environment. With RAG, summaries remain up-to-date concerning the regulatory requirements, an important requirement for compliance in such a fast-changing legal world. This will further allow cross-attention to be applied so that summaries can be personalised to correspond to the particular needs of varied users. Its unique

context makes the compliance process much more intuitive. This would aid better decision-making and reduce the number of ubiquitous non-compliances.

4. Discussion of Challenges

We identify several significant challenges in creating such a method for customisation, summarising regulatory compliance. One major issue is that legal language is complex: legal documents' sentences are complete with jargon and complex sentence structure [19]. In such cases, the system might need domain-specific language models and NLP techniques to cater for this. Extensive collections of documents from different regulatory bodies bring performance issues, and thus, they need to look for solutions such as using distributed programming and optimised indexing for processing documents. Another challenge in KGs and ontologies is to ensure data precision and consistency; any inaccuracies will degrade the quality of such summaries. These knowledge structures could be validated regularly and routinely checked for integrity automatically. Computational complexity is also critical because, given large amounts of data at stake, the processes associated with it are resource-intensive. These challenges can be mitigated using parallel processing, GPU acceleration, and model optimisation.

Lowering scalability will make it easier to evolve the system to new regulations as they come out without requiring manual updates as frequently. Incremental updates and continuous learning might be employed to avoid system performance degradation over time. Additionally, the KGs and ontologies continuously need to be updated to incorporate regulation changes that could be resolved via automated update mechanisms backed by expert reviews. The proposed solution will be scalable and reliable through optimisation strategies, continuous updates, and cutting-edge technologies to address these challenges. It will improve decision-making while reducing risks of non-compliance in actual-world applications.

5. Conclusion

In conclusion, integrating ontologies, KGs and RAG to provide an overall solution to aid organisations in regulation compliance provides a stable and feasible approach. Therefore, it will offer intelligent support to organisational efforts in regulatory monitoring activities. To successfully develop this approach, detailed experiments quantify how much it overcomes the limitations of relying on LLMs alone in the legal and regulatory domains. In addition, domain ontologies for regulatory compliance and knowledge graphs for different regulations will furnish the research community with reusable artefacts to be used in other decision support systems.

References

[1] R. Nair, K. Levacher, M. Stephenson, Towards automated extraction of business constraints from unstructured regulatory text, in: D. Zhao (Ed.), Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Santa Fe, New Mexico, 2018, pp. 157–160. URL: https://aclanthology.org/C18-2034.

[2] M. Hinterleitner, C. Knill, Y. Steinebach, The growth of policies, rules, and regulations: A review of the literature and research agenda, Regulation amp; Governance 18 (2023) 637–654. URL: http://dx.doi.org/10.1111/rego.12511. doi:10.1111/rego.12511.

[3] A. Agarwal, S. Xu, M. Grabmair, Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, 2022. URL: http://dx.doi.org/10.18653/v1/2022.findings-emnlp.134. doi:10.18653/v1/2022.findings-emnlp.134.

[4] W. Hua, Y. Zhang, Z. Chen, J. Li, M. Weber, Mixed-domain language modeling for processing long legal documents, in: Proceedings of the Natural Legal Language Processing Workshop 2023,

Association for Computational Linguistics, 2023. URL: http://dx.doi.org/10.18653/v1/2023.nllp-1.7. doi:10.18653/v1/2023.nllp-1.7.

[5] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, E. Rahm, Construction of knowledge graphs: State and challenges, 2023. URL: https://arxiv.org/abs/2302.11509. arXiv:2302.11509.

[6] W. Liang, P. D. Meo, Y. Tang, J. Zhu, A survey of multi-modal knowledge graphs: Technologies and trends, ACM Computing Surveys 56 (2024) 1–41. URL: http://dx.doi.org/10.1145/3656579. doi:10.1145/3656579.

[7] G. Wang, W. Li, E. Lai, J. Jiang, Katsum: Knowledge-aware abstractive text summarization, 2022. URL: https://arxiv.org/abs/2212.03371. arXiv:2212.03371.

[8] R. C. Barron, V. Grantcharov, S. Wanna, M. E. Eren, M. Bhattarai, N. Solovyev, G. Tompkins, C. Nicholas, K. Rasmussen, C. Matuszek, B. S. Alexandrov, Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization, 2024. URL: https://arxiv.org/abs/2410.02721. arXiv:2410.02721.

[9] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on rag meeting llms: Towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, volume 24 of *KDD '24*, ACM, 2024, p. 6491–6501. URL: http://dx.doi.org/10.1145/3637528.3671470. doi:10.1145/3637528.3671470.

[10] Y. Zhang, C. Liu, M. Liu, T. Liu, H. Lin, C.-B. Huang, L. Ning, Attention is all you need: utilizing attention in ai-enabled drug discovery, Briefings in Bioinformatics 25 (2023). URL: http://dx.doi.org/10.1093/bib/bbad467. doi:10.1093/bib/bbad467.

[11] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, S. Azam, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access (2023). URL: http://dx.doi.org/10.36227/techrxiv.24171183.v1. doi:10.36227/techrxiv.24171183.v1.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[13] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Long, Y. Cai, Z. Li, Z. Zhang, Y. Mo, J. Gu, R. Jiang, Y. Wei, C. Xie, Milvus: A purpose-built vector data management system, in: Proceedings of the 2021 International Conference on Management of Data, SIGMOD '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2614–2627. URL: https://doi.org/10.1145/3448016.3457550. doi:10.1145/3448016.3457550.

[14] J. J. Miller, Graph database applications and concepts with neo4j, in: Proceedings of the southern association for information systems conference, Atlanta, GA, USA, volume 2324, 2013, pp. 141–147.

[15] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, Advances in neural information processing systems 26 (2013).

[16] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, E. Choi, Factkg: Fact verification via reasoning on knowledge graphs, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023. URL: http://dx.doi.org/10.18653/v1/2023.acl-long.895. doi:10.18653/v1/2023.acl-long.895.

[17] J. Zhou, X. Han, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Gear: Graph-based evidence aggregating and reasoning for fact verification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019. URL: http://dx.doi.org/10.18653/v1/p19-1085. doi:10.18653/v1/p19-1085.

[18] F. Sovrano, M. Palmirani, F. Vitali, Legal Knowledge Extraction for Knowledge Graph Based Question-Answering, IOS Press, 2020. URL: http://dx.doi.org/10.3233/faia200858. doi:10.3233/faia200858.

[19] X. Yang, Z. Wang, Q. Wang, K. Wei, K. Zhang, J. Shi, Large language models for automated qamp;a involving legal documents: a survey on algorithms, frameworks and applications, Inter-

national Journal of Web Information Systems 20 (2024) 413–435. URL: http://dx.doi.org/10.1108/ijwis-12-2023-0256. doi:10.1108/ijwis-12-2023-0256.