Towards Building a Legal Virtual Assistant Based on Knowledge Graphs

Douglas Raevan Faisal[1], Fariz Darari[1], Berty Chrismartin Lumban Tobing[2] and On Lee[3]

[1]Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2]CATAPA, Jakarta, Indonesia
[3]GDP Labs, Jakarta, Indonesia

Abstract

Knowledge graphs (KGs) are gaining more and more attention nowadays due to their usefulness in many applications. Virtual assistants (VAs) can have a beneficial role in providing a conversational user interface (UI) for KGs. The legal domain touches a crucial aspect of our lives and yet access to legal professionals remains an obstacle. The use of KG-based VAs in the legal domain can help alleviate such a problem, democratizing legal knowledge to anyone, anywhere, and anytime. In this paper, we present our industrial experience in building a VA on top of a legal KG about laws related to labor. We discuss several use cases and lessons learned from what we have done. A demo video to give a glimpse of our KG-based VA is available online at https://youtu.be/yNOsXQaK89E.

Keywords

Laws, Virtual Assistant, Knowledge Graphs

1. Motivating Scenario

Human resource management (HRM) is the process of managing people: recruitment, compensation & retention, training & skill development, legal support, and more [1]. One of the duties of HR departments is to answer employee questions about labor-related stuff. Oftentimes, HR departments get overwhelmed because they should answer repetitive questions from employees, not to mention the laws and regulations regarding labor might change frequently. To solve these problems, CATAPA, an AI-based HRM platform, intends to provide (among others) knowledge services of laws related to labor, which should meet the following two major requirements. First, the services should capture the rich, semantic representation of legal knowledge. Second, the services should support the advanced retrieval of legal knowledge in a user-friendly, conversational manner without the need for technical expertise. The second criterion is particularly important as most CATAPA users come from non-CS background. In order to realize such knowledge services, we explore the use of KGs and VAs based on the KGs, which will be detailed in later sections of this paper.

Joint Proceedings of ISWC2022 Workshops: the International Workshop on Artificial Intelligence Technologies for Legal Documents (AI4LEGAL) and the International Workshop on Knowledge Graph Summarization (KGSum) (2022)

douglas.raevan11@ui.ac.id (D. R. Faisal); fariz@ui.ac.id (F. Darari); berty.c.l.tobing@catapa.com (B. C. L. Tobing); onlee@gdplabs.id (O. Lee)

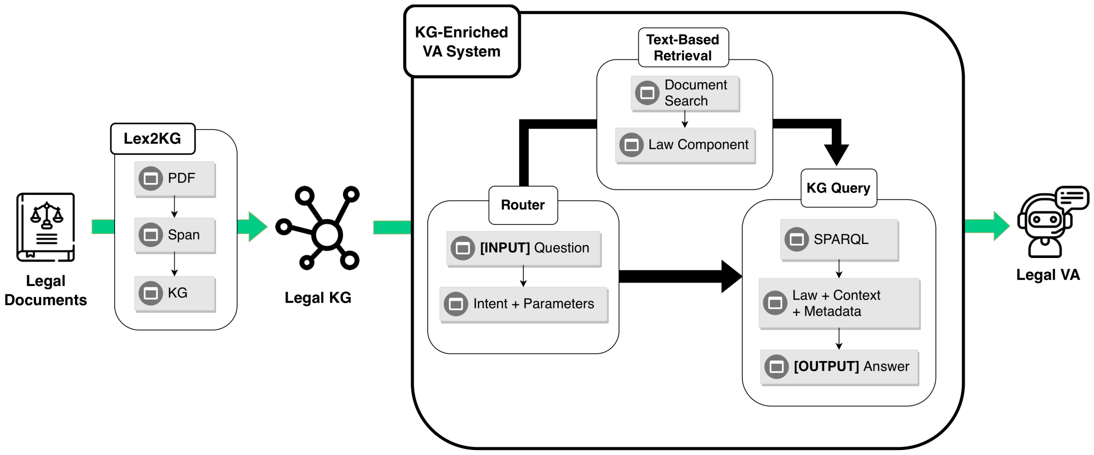

Figure 1: Architectural Flow of Lex2KG and KG-Based Legal VA

2. Architecture and Implementation

Knowledge graphs (KGs) are a way to represent knowledge by describing a collection of entities and linking them semantically based on their real-world relationships [2]. We base our approach to building a feature-rich legal VA on legal KGs built from law documents. In the following paragraphs, we elaborate on the architecture and implementation of our KG-based legal VA.

Architecture. The architecture comprises two separate systems as shown in Fig. 1. Since most laws are written in text documents, the first step is to convert those legal PDF documents into KGs, using the Lex2KG system [3]. Such conversion is driven by the Lex2KG ontology,[1] which contains legal-related classes (e.g., Legislation, Chapter, Article) and properties (e.g., cites, part of, amends). The second system is the legal VA that uses the generated KGs as the underlying source of data.

The legal VA receives user questions and then responds with law components that best answer the corresponding questions. The underlying process from input to output is realized through three main modules: router, KG querying (or KG-based retrieval), and text-based retrieval. The router takes and interprets user questions into the right business logic by calling the corresponding fulfillment function with detected intents and extracted information. Depending on the intent, answers to given questions are provided through either text-based retrieval or KG querying. Note that the resulting law component of text-based retrieval still has to be fed into KG querying in order to get the context and metadata (i.e., semantic enrichment) of the result.

Implementation. We employ several AI-related technologies to build a robust legal VA. Knowledge graphs (KGs) serve as one of the core technologies, allowing linked data representation of law documents. Such representation is capable of reconstructing the complex structure of law documents where each law component is interlinked to each other, e.g., law references/citations and law amendments.

[1] The ontology is available at https://bit.ly/lex2kg-o.

Aside from KGs, we also rely on a couple of technologies to build the legal VA: natural language processing (NLP) and text-based information retrieval (IR). NLP is leveraged to analyze user questions and detect intents (through text classification) as well as important parameters (through entity tagging) of these questions. On the other hand, text-based retrieval comes in handy for answering free-format legal questions [4]. We use the BM25 [5] text ranking system for its fast querying as well as adaptability to new data. Furthermore, we make use of question pattern matching similar to [6] in our IR services.

3. Use Cases

In this section, we discuss in four use cases the added value over question answering (QA) capabilities of our KG-based legal VA.

QA over Legal Structures. In most cases, users may encounter law references to be looked up (e.g., Article 8 of Act 11/2020 of the Republic of Indonesia) but with only access to the PDF documents. In that case, it will be difficult for them to skim pages to find a single article. Simply by utilizing the Lex2KG ontology [3], the legal VA is able to understand the underlying legal structure in the QA process (e.g., What is the content of Article X of Act Y of the Republic of Indonesia?). As such, users can instantly get the law components they need by simply mentioning the law reference to our VA system.

QA Integrating Multiple Components: Sanctions. In law enforcement, sanctions are laid out to reduce the risk of law violations. Figuring out the sanctions of a violation can be burdensome and that it requires two steps: (*i*) locating the violated law part, and (*ii*) finding the law part containing the corresponding sanctions. We need to ensure that users are able to find such information on the violated law part and its sanctions quickly.

Here we demonstrate one of the advantages of KGs in integrating multiple law components. First, the system finds the relevant law component that is being violated. Next, the system infers which sanction corresponds to the violation through links in KGs. Such an approach can be effective in the common case of users only describing the violation in order to get the relevant sanctions (e.g., What are the sanctions for the employment of underaged workers?).

QA over Legal Content with Context Recommendation & Enrichment. KGs are rich of metadata and contexts. In terms of legal VA, KGs can provide a variety of context enrichment during the answering process, such as law metadata, citations, and implementing regulations (by laws of lower-order). Through such information, users can get a comprehensive understanding of the laws and their parts.

QA over Laws with Revisions and Amendments. Besides metadata and contexts, laws and regulations are regularly updated through amendments. The nature of linked data in KGs allows for easier ontology modeling in handling revisions and amendments. By leveraging this advantage, we can make sure that the legal VA will always provide the laws currently in force (unless asked otherwise).

4. Lessons Learned

Throughout the development and exploration of our KG-based legal VA, we encounter a set of lessons learned, which can be valuable to anyone wanting to build a legal VA.

Evaluation over QA Dataset. In order to get an understanding as to how our legal VA system may perform in practice, we develop a dataset[2] of 200 question-answer pairs and test our system over the dataset. The dataset concerns the Indonesian Labor Law (Act 13/2003 as amended by Act 11/2020) and comprises four different question types: (*i*) definition lookup; (*ii*) law component lookup; (*iii*) sanctions; and (*iv*) domain knowledge.

The evaluation results are as follows. When we isolate the tests only for the KG part, all of the questions are answered correctly (100% pass rate). As for the end-to-end evaluation (which incorporates intent classification and parameter extraction), our system answers correctly 60% of all the questions, significantly higher than a plain BM25 retrieval system (13%).

Hybrid Approach to Legal VA. We identify the need for a hybrid approach to legal VA during the early stage of development. We incorporate NLP and IR techniques in tandem with KG technologies to fit the more flexible nature of user questions that align with our business use cases. Such a hybrid approach is also adopted by the Project Lynx [7, 8], which develops a legal KG for European Union (EU) laws. The project makes use of NLP techniques for the enrichment service and IR for the document management service.

Answers from Multiple Parts. In answering user questions, there are cases where a single law component is not sufficient. This issue is commonly known as multi-hop QA [9]. We tackle this by means of metadata and contexts from our KGs, which are particularly useful for the use cases of answering questions related to sanctions and law recommendations based on citation links.

We also identify an alternative that can be explored in this realm: using a refined ontology that can capture higher-level semantics rather than just legal structures. For example, when a user asks questions about labor, the legal VA can provide the definition of labor, the rights and obligations, and how labor is related to other similar entities (e.g., employers, companies). In related studies, SALKG [10] focuses on a semi-automatic approach to annotating high-quality semantics for constructing legal KGs, whereas Veena et al. [11] use refined legal ontologies to better provide information about legal penalties. The identification and linking of legal entities can also be improved through named-entity recognition (NER), as demonstrated in [12] on Greek legislations.

Completeness of Laws. For our business use case, we develop a domain-specific legal VA about the labor topic. As laws govern many aspects of citizen's lives, a legal VA aiming to tackle a more generic domain must ensure the completeness of the included laws. Accounting for large-scale law data may attribute to two issues: scalability and accuracy. In terms of scalability, the system architecture must scale well with the large variety of topics and the addition of law

[2] Our dataset is available at https://s.id/IDLaborLawVATestCases.

documents over time. In terms of accuracy, increasing the scope and depth of laws may risk answer overloading: too many seemingly relevant answers can be returned and that seeking the right answer can be more challenging.

Generalization to Other Law Types and Jurisdictions. We explore legal VA approaches under the Indonesian jurisdictions. Thus, we have to deal with the different law types that exist in Indonesia, e.g., Laws/Acts (*Undang-Undang*), Government Regulations (*Peraturan Pemerintah*), and Local Regulations (*Peraturan Daerah*). Under different jurisdictions, laws may apply differently. For example, in the EU, the legal KG must consider the different languages used by different EU countries [7]. Zhong et al. [13] argue that one of the challenges in generalizing legal KGs (and their applications) is that the same legal concepts may have different representations and meanings under different jurisdictions.

5. Conclusions

In line with the goal of CATAPA as an AI-based human resource management (HRM) platform provider, we have developed a legal virtual assistant (VA) to make legal knowledge about labor more accessible for employees as well as HR practitioners. We have explored the use of knowledge graphs (KGs) as the foundation of our VA. The major requirements of capturing the semantic representation of legal knowledge and providing a user-friendly and conversational interface have been fulfilled with the KG-based legal VA. We have also touched on the lessons learned during the development and exploration process. This set of lessons learned can provide useful insights for practitioners and researchers in building KG-based legal VAs.

Acknowledgments

We are grateful to Stefanie Suanita (CEO of CATAPA) for her support of this research work.

References

[1] Human Resource Management, University of Minnesota Libraries, Minneapolis, MN, 2016. URL: https://open.lib.umn.edu/humanresourcemanagement/.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, Morgan & Claypool, 2021. URL: https://kgbook.org/.

[3] M. Abdurahman, F. Darari, H. Lesmana, M. Hartopo, I. Rhesa, B. C. L. Tobing, Lex2KG: Automatic Conversion of Legal Documents to Knowledge Graph, in: ICACSIS, 2021.

[4] S. Khazaeli, J. Punuru, C. Morris, S. Sharma, B. Staub, M. Cole, S. Chiu-Webster, D. Sakalley, A Free Format Legal Question Answering System, in: NLLP, 2021.

[5] K. S. Jones, S. Walker, S. E. Robertson, A probabilistic model of information retrieval: development and comparative experiments: Part 2, Inf. Processing & Management 36 (2000) 809–840.

[6] A. A. Zulen, A. Purwarianti, Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question, in: PACLIC, 2011.

[7] V. Rodriguez-Doncel, E. Montiel-Ponsoda, Lynx: Towards a legal knowledge graph for multilingual Europe, Law in Context 37 (2020) 175–178.

[8] J. M. Schneider, G. Rehm, E. Montiel-Ponsoda, V. Rodríguez-Doncel, P. Martín-Chozas, M. Navas-Loro, M. Kaltenböck, A. Revenko, S. Karampatakis, C. Sageder, J. Gracia, F. Maganza, I. Kernerman, D. Lonke, A. Lagzdins, J. B. Gil, P. Verhoeven, E. G. Diaz, P. B. Ballesteros, Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain, Information Systems 106 (2022).

[9] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: EMNLP, 2018.

[10] M. Tang, C. Su, H. Chen, J. Qu, J. Ding, SALKG: A Semantic Annotation System for Building a High-quality Legal Knowledge Graph, in: IEEE International Conference on Big Data, 2020.

[11] G. Veena, D. Gupta, A. Anil, S. Akhil, An Ontology Driven Question Answering System for Legal Documents, in: ICICICT, 2019.

[12] I. Angelidis, I. Chalkidis, M. Koubarakis, Named Entity Recognition, Linking and Generation for Greek Legislation., in: JURIX, 2018.

[13] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence, in: ACL, 2020.