Collaborative System for Question Answering in German Case Law Documents

Christoph Hoppe, Nico Migenda, David Pelkmann, Daniel Hötte, Wolfram Schenck

Collaborative System for Question Answering in German Case Law Documents

Christoph Hoppe[1], Nico Migenda[1], David Pelkmann[1], Daniel Hötte[2], and Wolfram Schenck[1]

[1] Center for Applied Data Science, Bielefeld University of Applied Sciences/Gütersloh, Germany
[2] Faculty of Business, Bielefeld University of Applied Sciences/Bielefeld, Germany
{christoph.hoppe, nico.migenda, david.pelkman, daniel.hoette, wolfram.schenck}@fh-bielefeld.de

Abstract. Legal systems form the foundation of democratic states. Neverthe- less, it is nearly impossible for individuals to extract specific information from comprehensive legal documents. We present a human-centered and AI-supported system for semantic question answering (QA) in the German legal domain. Our system is built on top of human collaboration and natural language processing (NLP) -based legal information retrieval. Laypersons and legal professionals re- ceive information supporting their research and decision-making by collaborating with the system and its underlying AI methods to enable a smarter society. The internal AI is based on state-of-the-art methods evaluating complex search terms, considering words and phrases specific to German law. Subsequently, relevant documents or answers are ranked and graphically presented to the human. In ad-dition to the novel system, we publish the first annotated data set for QA in the German legal domain. The experimental results indicate that our semantic QA workflow outperforms existing approaches.

Keywords: Question Answering · Information Retrieval · Human-AI interface design · AI-supported decision making · Legal Research

1 Introduction

Legal systems are an indispensable pillar in constitutional nations [17]. With the ongoing digitization of the legal system, legal documents are increasingly stored digitally. The question arises of how a legal layperson is supposed to gather information from the overwhelming number of court documents. To extract legal information efficiently a

collaborative system is necessary [7]. Traditionally, this meant consulting a specialist, which is neither cheap nor fast. With the rise of large-scale language models and question answering (QA) new opportunities to assist legal laypersons are emerging. Current legal Information Retrieval (IR) systems, databases, and commercial search engines store and process large amounts of legal documents. Unfortunately, extracting a specific passage or answer from a corpus of documents to a posed legal question is either highly time-consuming or impossible, as existing approaches do not provide capabilities for modern QA and semantic search. Moreover, actual research in legal IR and QA shows a lack of sufficient language models and data sets in the German language [22]. This makes it necessary to develop a QA system in the German legal domain that can return a precise answer to a posed question in a corpus of large legal documents, providing collaborative and transparent access to legal information [25].

In this paper, we present a system for semantic QA in German case law documents that follows the guidelines of state-of-the-art search engines. Thereby, our system focuses on human-centered AI collaboration that enables efficient passage retrieval and QA across large-scale legal documents by interacting with humans and autonomously providing suggested answers to legal search queries. We use different retrieval methods as well as a self-trained reader model based on Efficiently Learning an Encoder that Classifies Token Replacement Accurately (ELECTRA) [5]. In order to train and evaluate the presented language models in the field of law, we created a hand-annotated data set consisting of 226 question-answer pairs from German case law documents. In addition, we evaluate multiple end-to-end semantic search workflows consisting of different retriever-reader combinations in comparison to our model. In doing so, we show that fine-tuning a pre-trained language model to a specific domain shows great results even with a small amount of labeled data. Thus, our research bridges the gap between existing legal IR systems and modern human-centered AI technologies. The intelligent system further contributes towards a smarter society. Our approach supports both le- gal laypersons who need an initial assessment of their problems and legal professionals such as lawyers, who need assistance at their research activities.

2 Related Work

Legal IR systems and databases were originally built to store and retrieve large amounts of various legal documents. *CourtListener* and the *Caselaw Access Project (CAP)* are initial approaches that aim to provide free access to published court decisions of the United States legal system in a uniform format enriched with additional metadata [9, 14]. The two most popular and up-to-date databases for legal research in German language are *Beck Online* and *JURIS*, which publish large collections of court decisions as well as excerpts from legal texts and legal commentaries [1, 8]. The challenge of offering legal documents in

machine-readable formats is solved by the legal IR systems *Openlegaldata* and the *Finlex data bank*. Both provide free and public access to legal documents in various data formats and contribute to further machine processing and data analysis in the legal field [15, 18]. However, traditional IR systems are limited inthe retrieval of specific passages or answers to a posed question, resulting in imprecise and inefficient results. This has motivated the scientific community to introduce methods that provide QA possibilities for a large amount of legal documents. The majority of recent research in the field of QA generally follows the paradigm of a retriever- reader based search workflow which selects relevant documents and extracts a specific answer to a posed question [4]. Deep learning and large language models represent the given documents as dense vectors, taking their semantics and surrounding context intoaccount. Especially the deep representation of documents [6, 20, 23] leads to first le- gal QA approaches in the English and Chinese language [2]. Moreover, ontology and knowledge graph-driven methods have been applied in the field of Arabic and Chinesejurisprudence [24, 12, 11].

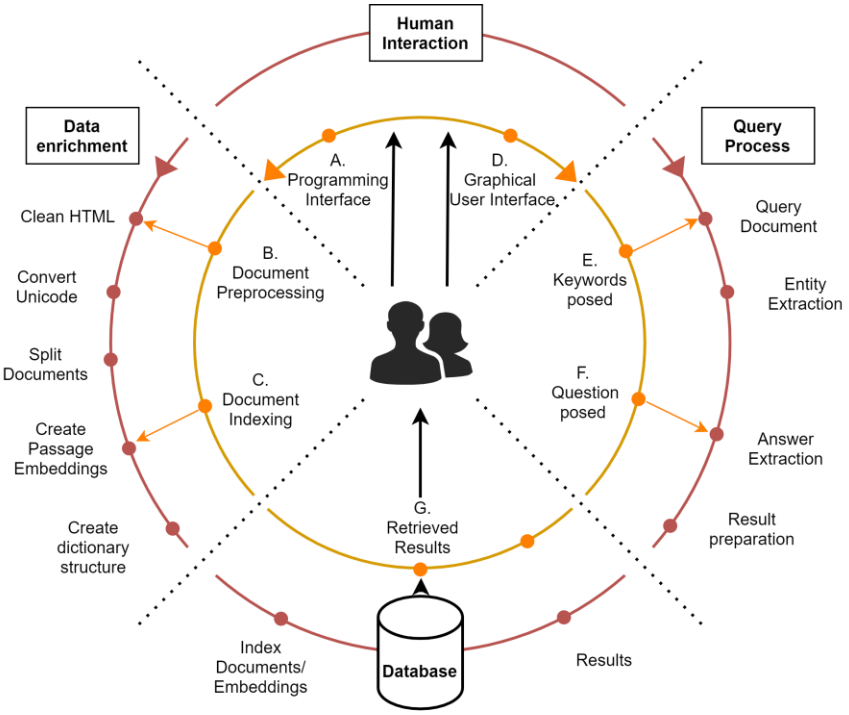

Fig. 1. Presentation of the collaborative QA system. The inner (orange) circle represents the hu-man interaction with the system to either enrich (counterclockwise) or to query (clockwise) thedatabase. The outer (red) circle represents the AI-toolchain to process the human input. Both circles are deeply interlinked.

3 Human-AI Collaborative QA System

Human-AI collaboration aims to accomplishing a shared goal by deep interaction between humans and AI. Legal laypersons alone are not capable of extracting information from large numbers of law documents. Instead of searching through many documents, the human interacts via a graphical interface with an AI that performs the document search and QA tasks. The AI itself consists of a toolchain of exchangeable statistical and machine learning components to solve the search tasks more efficiently. The system (Fig. 1) consists of two connected and continuously interaction layers: the human layer(front end) and the AI layer (back end).

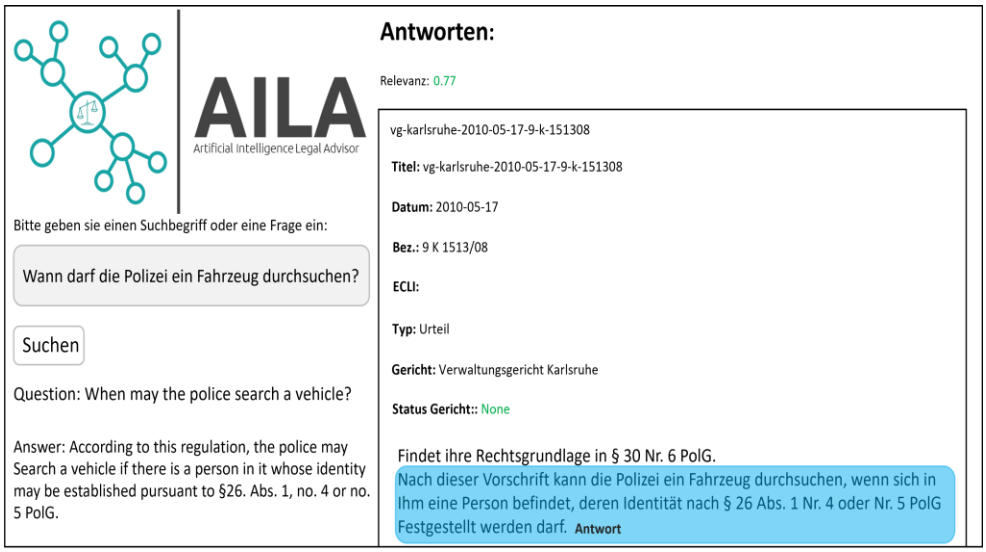

Fig. 2. Section of the graphical user interface. Users can interact with the system and submit a question to the interface (left). Afterwards, they receive back a concrete answer extracted from a legal document (right), enriched with a relevance score and additional metadata.

3.1 Human Interaction

Our system focuses on collaborative interaction between humans and the QA system. For this purpose, humans can engage with various interfaces in order to perform operations on the system (Fig. 1, inner circle). In general, there are two different ways to interact with the system. Initially, it is possible to interact with the AI using the programming interface

(Fig. 1, A.) to enhance the knowledge base by inserting new legal documents. These documents are automatically pre-processed, embedded, indexed and saved in the database (no actual programming skills are required). Secondly, a search query can be submitted using the graphical human-AI interface (Fig. 1, D.). This query can either consist of individual keywords or contain a specific legal question (Fig. 2). These queries can contain law specific terms (e.g., the § symbol for a paragraph). De- pending on the search term, different components of the AI toolchain, which are de- scribed in the following section 3.2, are used to find the best matching documents. The top results are then prepared in such a way that the human can directly extract the requested information. In an evaluation loop the presented documents are optionally rated, to enhance the quality of future requests.

3.2 AI Layer

The AI layer is the collection of exchangeable AI components located in the back end of the system (Fig. 1, outer circle). When the human submits a search term, the query process is started via the graphical interface. This triggers the process of retrieving relevant information related to a search query. In a first step, a combination of statistical methods and deep sentence transformer models are used for the task of document retrieval (Fig. 1, E. Keywords posed). To find the most relevant passages, the model transforms a posed question or keywords into a vector of the dimension $d = 768$. This computed vector is mapped to a shared embedding space. Subsequently, the cosine similarity is used to compute the distances between the embedded question or keywords and stored passages. Afterwards, the next $top_k = 10$ relevant passages related to the search re- quest are retrieved [20, 21]. If only single keywords are passed to the system, the retrieved passages are checked for legal entities using a BERT model [6] trained for the task of named entity recognition [13]. Whenever a concrete legal question is passed to the system, the answer extraction method is additionally integrated into the process (Fig. 1, F. Question posed). Therefore, an ELECTRA model is added to the process as a reader. This model receives the relevant passages returned by the retriever as well as the questions posed. In the following, the model extracts the exact answers from the given passages. In our approach, given the $top_k = 10$ passages returned by the retriever, the reader model is advised to extract the $top_k = 5$ answers that are most relevant with respect to the posed question. Finally, the passages or answers found are ranked according to relevance and presented to the human on the graphical human-AI interface.

When the AI is called using the programming interface to enhance the knowledge base, different toolchain components are used. The main tasks of the programming interface are the execution of the document pre-processing and document indexing processes. Their general function is to prepare the raw legal documents within our data set and to store the pre-processed documents inside the database. The document pre-processing (Fig. 1, B.

Document preprocessing) includes various methods such as the removal of HTML elements from the plain text and the conversion of Unicode symbols. In order torespect the maximum processable token length of different retrieval models and to im- prove the performance of the retrieval process, long documents are split into passages of 200 words each. Using the indexing method, both the plain text and the metadata are stored in the database (Fig. 1, C. Document indexing). Moreover, we generate deep vector representations of the passages with a dimension of $d = 768$ in order to perform a semantic search and QA [20]. These vectors are also stored in the database that acts as a shared embedding space. Once all the texts, metadata and passage vectors in the database have been successfully indexed, the indexing process is complete.

4 Experiments and Results

The experiments compare the performance of our system with different underlying models. We test several pre-trained models as well as s self-trained reader model. More- over, we introduce the self-annotated data set LegalQuAD for QA tasks in German case law documents.

4.1 Creation of the LegalQuAD Data Set

To evaluate or fine-tune retriever and reader models for the task of QA, annotated datasets consisting of question-answer pairs are necessary. To the best of our knowledge, there is currently no annotated data set published for QA in German legal documents. To overcome this problem, we created a hand-annotated LegalQuAD for training and evaluation purposes. The entire data set consists of 226 question-answer pairs from Ger- man case law documents of various legal fields and is structured in the SQuAD format[19, 16]. The data annotation itself was performed by lawyers who are familiar with NLP and have received intensive training on the data annotation process. During the annotation phase, various passages from German case law documents were presented to the lawyers. While reading a passage, the lawyers were asked to formulate a specificquestion regarding the given passage and highlight a corresponding answer. To ensure the creation of a diversified data set, both complex questions that need to be answered over a span of several sentences as well as shorter questions, that can be answered in a few words are formulated. In addition, the annotators were instructed to rephrase posed questions with synonyms to avoid lexical overlap between question-answer pairs. Fur thermore, it was reviewed that all formulated questions are self-sufficient and can be answered completely with the knowledge indicated in the respective text.

4.2 Model Training and Evaluation Metrics

We compared the performance of different combinations of the entire QA workflow introduced in Section 3 on our annotated LegalQuAD. Therefore, we selected several publicly available models and tested their performance on our LegalQuAD test data set (see Table 1). Given the results of previous studies, we choose to compare the re- trieval methods *BM25* and *MFAQ* in combination with the reader models *GELECTRA- base-GermanQuAD* and *GELECTRA-large-GermanQuAD* [10]. Moreover, we trained our own reader model *GELECTRA-large-GermanQuAD-LegalQuAD* by fine-tuning the pre-trained model *GELECTRA-large-GermanQuAD* [16, 3]. For this purpose, we created a training data set by selecting 200 random question-answer pairs from LegalQuAD. Afterwards, we trained the model for two epochs with a learning rate of $l = 1e - 5$ using Adam as an optimizer and a batch size of $b = 10$ as well as a maximum sequence length of $sl = 256$ tokens. To evaluate the implemented models, we consider several evaluation metrics that show whether the answer span predicted by the model matches the correct answer:

Exact Match (EM) is a metric that measures the proportion of documents where the predicted answer span exactly matches the correct answer span. This metric is very precise and restrictive. For example, a predicted answer *A: "§ 15 BGB."* would result in a score of zero if the answer labeled as correct was *A: "In § 15 BGB."* because this answer span does not match the expected answer exactly. A metric that measures the ratio of over- lapping words between the labeled and predicted answer span is the *F1-score*. Thus, this metric is more forgiving than the EM and closer to a human opinion regarding the similarity of two predicted answers [16].

4.3 Results and Discussion

The results of our experiments are presented in Table 1. On the one hand, it has been shown that the combination of the retriever models *MFAQ* and *BM25* in combination with our self-trained reader *GELECTRA-large-GermanQuAD-LegalQuAD* shows the best performance in respect to the EM and F1-scores. On the other hand, it can be observed that the results of pre-trained models is significantly weaker in comparison to our fine-tuned approach.

The results of the experiments indicate that our presented workflow towards a collabo rative QA system is functional and able to achieve valuable results compared to state-of-the-art approaches. In particular, it was demonstrated that the human-AI-collaboration regarding the preprocessing and indexing of German legal documents, as well as the query process, is straightforward to manage for both legal laypersons and lawyers, with-out the need of programming knowledge. In comparison to our fine-tuned model, the pre-trained reader models show weak scores especially in the EM evaluation metric. We argue that the significant increase in the EM score after the fine-tuning process is a result of the fact that law has a very strong and complex domain language which is hard to

generalize by pre-trained models.

Table 1. Performance of the QA workflow consisting of retriever and reader models on the LegalQuAD test data set. The results shown were calculated for retriever $top_k = 10$ and reader $top_k = 5$. Model type and training dataset are included in the model name.

Retriever	Reader	Exact Match	F1-Score
BM25	GELECTRA-base-GermanQuAD	0.083	0.58
BM25	GELECTRA-large-GermanQuAD	0.12	0.67
BM25	GELECTRA-large-GermanQuAD-LegalQuAD	0.82	0.98
MFAQ (Emb)	GELECTRA-base-GermanQuAD	0.083	0.48
MFAQ (Emb)	GELECTRA-large-GermanQuAD	0.083	0.51
MFAQ (Emb)	GELECTRA-large-GermanQuAD-LegalQuAD	0.5	0.72
MFAQ (Emb) + BM25	GELECTRA-large-GermanQuAD-LegalQuAD	**0.83**	**0.98**

The fact that our reader models show a significant increase in the performance compared to the publicly available models suggests that our research can be beneficial in extracting specific answers from large legal documents in the future. Furthermore, it could be shown that the fine-tuning process of a pre-trained language model can be successfully performed even with a small amount of labeled data. Thus, there is great potential for adapting existing models to a specific domain. As the field of IR moves from statistical document retrieval to AI-based methods, there is a growing need to develop collaborative systems for AI-supported QA and semantic search in the legal domain. With the increasing use of AI, it becomes especially important to bring humans into the center of the process chain and make legal information accessible to both legal laypersons and legal professionals through a human-AI inter- face. The research results presented will enable us to improve existing IR systems from the legal sector like *Beck Online* and *JURIS* and enhance their process with QA functionalities. In doing so, we create transparent and straightforward search capabilities in comprehensive legal documents for all humans, contributing towards a smarter society.

5 Conclusion and Future work

In this paper, we introduced our approach for a collaborative QA system in German case law documents. We have shown that retrieving information from legal documents and performing extractive QA in the field of German case law are relevant and un- solved problems in the legal system, which can be solved by our semantic search and QA approach. Moreover, we established a human-AI interface and described its applied methods, models and human interactions in detail. To encourage further research in the

area of QA in the German language, we hand-annotated and published the data set LegalQuAD, consisting of question-answer pairs derived from German case law documents. To the best of our knowledge, LegalQuAD is the first annotated QA data set forthe training and evaluation of legal language models in the German language. Based on this published data set, we evaluated and trained models for the task of QA in German case law documents. Our experiments show that our workflow leads to significantly better results in terms of the EM and F1-score than previously published approaches. Addressing the time-consuming research activities of lawyers, our contribution aimsto bridge the gap between modern language technologies and the traditional-oriented field of law. Rationalizing inefficient information seeking for legal professionals allowthem to concentrate their work on the actual decision-making processes and find fasterconformity of the given cases with specific legal indications. This kind of efficiency may lead to an improvement in the performance of legal professionals which result in a higher quality of their consultation for individuals or entities. Furthermore, we are highly committed to reducing the barriers to society's access to legal information and empowering a smart society to obtain legal information independently from authorities.For this reason, we have ensured that our collaborative QA system can also be used bylegal laypersons and enables transparent and straightforward search capabilities in largelegal document collections. The code and data set related to the paper can be found at *https://www.github.com/Christoph911/Pro-Ve_2022_Appendix*

Acknowledgements We would like to thank Nicole Salvi for her support of the project by contributing to the annotation of the published dataset and the literature review conducted, as well as providing her excellent knowledge of the legal sector.

References

1. Beck, V.C.: Beck-online - die datenbank. https://beck-online.beck.de, accessed: 2022-02-17
2. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-BERT: The muppets straight out of law school. Findings of the Association for Computational Linguistics (EMNLP) pp. 2898–2904 (2020)
3. Chan, B., Schweter, S., Möller, T.: German's next language model. Proceedings of the 28th International Conference on Computational Linguistics (ACL) pp. 6788–6796 (2020)
4. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading wikipedia to answer open-domain questions. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) pp. 1870–1879 (2017)
5. Clark, K., Luong, M.T., Brain, G., Brain, Q.V.L.G., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. Proceedings of the 8th International Conference on Learning Representations (ICLR) (2020)
6. Devlin, J., Chang, M.W., Lee, K., Google, K.T., Language, A.I.: BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies (NAACL) pp. 4171–4186 (2019)
7. Enders, P.: Einsatz künstlicher intelligenz bei juristischer entscheidungsfindung. Juristische Arbeitsblätter pp. 721–735 (2018)
8. juris GmbH: Uris - das rechtsportal. https://juris.de/jportal/nav/index. jsp/ (2022), accessed: 2022-02-17
9. Harvard-Law-School: Caselaw access project. https://lil.law.harvard.edu/ projects/caselaw-access-project/(2013), accessed: 2022-03-05
10. Hoppe, C., Pelkmann, D., Migenda, N., Hötte, D., Schenck, W.: Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. Proceedings of the 4th Artificial Intelligence and Knowledge Engineering Conf. (AIKE) (2021)
11. Huang, W., Jiang, J., Qu, Q., Yang, M.: Aila: A question answering system in the legal domain. Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI) pp. 5258–5260 (2020)
12. Kourtin, I., Mbarki, S., Mouloudi, A.: A legal question answering ontology-based system. Proceedings of the 14th International NooJ Conference pp. 218–229 (2021)
13. Leitner, E., Rehm, G., Moreno-Schneider, J.: A dataset of German legal documents for named entity recognition. Proceedings of the 12th Language Resources and Evaluation Con- ference (LREC) pp. 4478–4485 (2020)
14. Lissner, M.: Courtlistener: A platform for researching and staying abreast of the latest in the law. Master thesis (2010)
15. Ministry-Justice-Finland: Finlex data bank. https://finlex.fi (2016), accessed:2022-03-05
16. Möller, T., Risch, J., Pietsch, M.: GermanQuAD and germanDPR: Improving non-english question answering and passage retrieval. Proceedings of the 3rd Workshop on Machine Reading for Question Answering (MRQA) pp. 42–50 (2021)
17. van Opijnen, M., Santos, C.: On the concept of relevance in legal information retrieval. Artificial Intelligence and Law **25**, 65–87 (2017)
18. Ostendorff, M., Blume, T., Ostendorff, S.: Towards an open platform for legal information. Proc. of the ACM/IEEE Joint Conference on Digital Libraries (JCDL) pp. 385–388 (2020)
19. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 2383–2392 (2016)
20. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP) pp. 3982–3992 (2019)
21. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval **3**, 333–389 (2009)
22. Sugathadasa, K., Ayesha, B., de Silva, N., Perera, A.S., Jayawardana, V., Lakmal, D., Perera, M.: Legal document retrieval using document vector embeddings and deep learning. Ad- vances in Intelligent Systems and Computing - Intelligent Computing **857**, 160–175 (2018)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I.: Attention is all you need. Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS) pp. 5999–6009 (2017)
24. Veena, G., Gupta, D., Anil, A., Akhil, S.: An ontology driven question answering system for legal documents. Proceedings of the 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) pp. 947–951 (2019)
25. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does nlp benefit legal system: A summary of legal artificial intelligence. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) pp. 5218–5230 (2020)