International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 6, December 2021, pp. 5450~5457
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i6.pp5450-5457

An overview of information extraction techniques for legal document analysis and processing

Ashwini V. Zadgaonkar[1], Avinash J. Agrawal[2]
[1,2]Department of Information Technology, Shri Ramdeobaba College of Engineering and Management, Nagpur, India
[2]Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Article Info

Article history:

Received Jan 2, 2021
Revised Apr 17, 2021
Accepted May 11, 2021

Keywords:

Deep learning
Information extraction
Information retrieval
Knowledge base population
Legal text processing

ABSTRACT

In an Indian law system, different courts publish their legal proceedings every month for future reference of legal experts and common people. Extensive manual labor and time are required to analyze and process the information stored in these lengthy complex legal documents. Automatic legal document processing is the solution to overcome drawbacks of manual processing and will be very helpful to the common man for a better understanding of a legal domain. In this paper, we are exploring the recent advances in the field of legal text processing and provide a comparative analysis of approaches used for it. In this work, we have divided the approaches into three classes NLP based, deep learning-based and, KBP based approaches. We have put special emphasis on the KBP approach as we strongly believe that this approach can handle the complexities of the legal domain well. We finally discuss some of the possible future research directions for legal document analysis and processing.

Corresponding Author:

Ashwini V. Zadgaonkar
Department of Information Technology
Shri Ramdeobaba College of Engineering and Management
Nagpur, 440013, India
Email: ashwinizadgaonkar24@gmail.com

1. INTRODUCTION

Nowadays a lot of information is available on the internet in a structured and unstructured form stored in multiple documents. This information belongs to different domains and needs to be analyzed and processed to extract the desired piece of information for a particular task. Manual processing and analysis of such a large repository of documents demand too much efforts and it will be very much time consuming also. To overcome these problems, automatic information processing and analysis is the need of the hour. Information retrieval and information extraction are the tasks required for automatic document analysis. Information extraction deals with automatically extracting relevant information for a particular application problem from the available corpus and represents it in a structured machine-readable format. Information retrieval gets relevant information sources whereas information extraction automatically extracts relevant information from those sources in a structured format. To differentiate between information retrieval (IR) and information extraction (IE) one can say that IR is a task that will locate the desired document form a large collection whereas IE focuses on extracting the exact piece of information from a document to solve user query. Generally, IE processes human language texts employing natural language processing (NLP) techniques. Automatic document analysis is desired by different domains like biomedical, administration, financial, literature, journalism, and many more. Researchers all over the world are using a combination of

different AI techniques such as natural language processing and understanding, named entity recognition, Relation extractions, Semantic role labeling, dependency parsing, and various machine learning models of classification to design and implement automatic document analysis systems.

2. LEGAL DOCUMENT ANALYSIS AUTOMATION

A legal domain expert can extract desired information from lengthy legal documents irrespective of its type, structuring, layout as per one's domain expertise but it is a time consuming and error-prone process. So automatic information extraction from legal documents is highly desired. Information extraction from legal documents will be directed by individual business requirements. The extracted information can be: i) stored in databases for future references, ii) for analysis and decision-making mechanism, iii) as an input to some other legal understanding task.

2.1. Need for legal document automation

Legal documents exist in different forms like legal contracts, law commission reports, tribunals, case judgments, different acts, online contracts, and many more forms. In India, different courts publish their legal proceedings every month for future reference of people. The number of manual efforts and time required to process these heterogeneous, unstructured, voluminous legal documents is too much. So there is a need for automation for analyzing legal documents from Indian law systems beneficial for legal practitioners as well as a common man for a better understanding of the legal domain.

Whenever a legal expert prepares a case file for a given case then it becomes mandatory to refer to the previous judgments given by different courts to build a strong case foundation. Manually going through thousands of declared court judgments and identifying the relevant information for the case in hand is very much time consuming and labor-intensive task. In today's internet era, there are some good search engines available to effective search in the legal domain. In discussion with legal experts, we come to know that legal search is generally driven by the following parameters, as shown in Figure 1.

a. Specific act, section, article, rule or an order of an act: while preparing for a case a lawyer need to refer to a particular act no or article no and wants to extract judgements given based on these acts or article no.
b. Appellant/respondent name: there were some landmark judgements exists in legal domain which are well known by appellant names. So legal expert may want to search such relevant cases.
c. Subjects/facts/phrases/keywords: legal search based on some facts or keywords or domain phases are highly desired in legal domain.
d. Judges names: There are some millstone cases well known for its judgements given by famous judges which is repetitively refer by lawyers for a relevant domain.
e. Head-note of the judgments: head notes of a particular case acts as a summary and generally manually created. So while preparing case notes browsing through head notes will help legal practitioners to identify relevant cases.
f. Time frame: legal expert try to identify recent relevant cases to prepare impressive case notes for which searching based on case year in important.

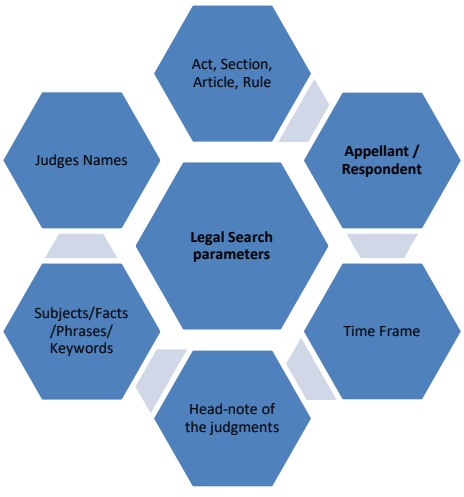

Figure 1. Different legal search parameters

2.2. How legal text differs?

Automatic legal document processing systems must understand some peculiar characteristics of domain corpus before further processing. Every year legal institutions produce thousands of documents in the form of legal contracts, law commission reports, tribunal, case judgments, acts, online contracts, citations. In countries like India, the Supreme Court of India, different state high courts, hundreds of district courts publish the legal proceedings in the public domain every month. But this large volume of publicly available legal data is not processed effectively to provide legal information to common people. One of the main reasons behind this is a complicated structure and lack of knowledge about legal language by common people. Some of the distinguishing features of a legal text in comparison to other domain texts are as shown in below.

a. Legal documents are too long as compared to documents in other domains.
b. Legal documents are having a complex internal structure containing a description of different acts, citations, and hierarchical form.
c. The vocabulary of legal documents consists of several domain-specific terminologies that may not be familiar with the non-legal community.
d. Ambiguity does exist in legal documents in the form of the different interpretations of the same content depending on the hierarchy of different courts, judges, or lawyers.

Citations are very important in the legal domain as compared to other domains and highlights of that particular case. The legal domain is quite promising for information retrieval and information extraction due to the large available corpus. As legal domain documents follow a peculiar layout, NLP techniques can process it better than extremely informal news and social media text. Hence, a knowledge base for automatically managing legal documents will be helpful for all types of users.

3. LITERATURE REVIEW

The approaches used for IE from legal documents are broadly classified into three categories and different legal document processing systems developed using these approaches are discussed below.

3.1. NLP techniques for legal text processing

By combining the power of artificial intelligence and computational linguistics, natural language processing (NLP) techniques help machines to "read" text by simulating the human ability to understand language. Some of the applications developed using NLP techniques are machine translation, automatic summarization, sentimental analysis, text classification, question answering. NLP represents the automatic handling of natural human languages like speech or text. The Law domain can be represented as a combination of language, logic, and conceptual relationships, and their analysis [1]. So, there is a wide scope of applying NLP techniques for legal information mining.

Kanapala *et al.* [2] provided a survey of different text summarization techniques recently used for legal text processing. This survey focuses on single as well as multiple document summarization techniques. These techniques were tested on different datasets like AustLII, HOLJ, Federal Court of Canada judgments. The techniques surveyed in this paper are divided into four categories namely the Linguistic Feature-based approach, Graph-based approach, Semantic role labeling based approach, and Classification based approach. Padayachy *et al.* [3] proposed an approach to design a comprehensive model to assist legal researchers in accessing legal data for the most applied case. The proposed approach is implemented using LegalCo. The legal database is provided by the organization. The proposed system is composed of four different modules namely information retrieval where query-dependent ranking and retrieval of the document is performed using the VSM model followed by the information extraction module which extracts the facts using NLP techniques for named entity recognition, relation extraction, and event extraction, the extracted facts are stored in graph database as labeled property graph (using Neo4j python library). The last module will return the recommendations in the form of the most applied case by doing a Query-independent ranking of obtained results.

Surdeanu *et al.* [4] proposed a method for extracting text relevant to litigation claims and entity mentions in each claim from hierarchical annotated legal domain data. They adopted a semi-supervised bottom-up approach for building a joint hierarchical conditional random field model using a combination of pseudo-likelihood and Gibbs sampling and proved that these models perform better in comparison with model adapting top-down approach. Constantino *et al.* [5] proposes a CLIEL system for annotating legal documents using XML tags to facilitate IE of data point instances such as date of the document, name of the party, governing law, and many more. The system is tested on the Set of 97 digitized commercial law documents of different formats, structures, and layouts. CLIEL system is using NLP techniques, java annotation pattern engine (JAPE), rule-based layout detection tree (RLDT) for information extraction from

Annotated XML document generated for each commercial law document and store it in a database for future reference.

María *et al.* [6] present an approach focused on validating and improving the quality of the results of an IE system based on the use of ontology that store domain knowledge. The proposed approach works on the output produced by the AIS system, an IE system specialized in analyzing Spanish legal documents. This approach is using Ontology specially designed for the legal domain and data curation process to validate the results obtained from *AIS* and store for future reference through the entity aligner module. Bommarito *et al.* [7] developed *LexNLP,* a Python package for extracting information from for legal and regulatory text. The objective behind the development is to support academic research as well as industrial applications. It is developed using NLP techniques and machine learning mechanisms to provide features like legal document segmentation, extracting structured info from the text, NER, converting text into feature vectors for the machine learning model. The model is built from real documents from SEC EDGAR and is open source. Savelka *et al* [8] proposed a framework for extracting important sentences from court judgments so that users need not refer to lengthy case documents for understanding statutory terms. They adopted techniques like measuring similarity among the case sentences and user queries, using the context model for sentences, query optimizations, and identify novel sentences for user queries. The proposed framework is tested on the labeled dataset of 4,635 sentences for three statutory queries. Kumar *et al.* [9] worked on finding similarity among the court judgments by using IR techniques and search engine mechanism. They have compared all term versus legal term cosine similarity method to prove that the legal term cosine similarity method performs better.

3.2. Deep learning techniques for legal text processing

Recently, Goodfellow *et al.* [10] becomes the popular choice of researchers for handling the complex and heterogeneous legal domain documents. Goldberg [11] provides an efficient approach to outperform traditional rule-based, dictionary-based, and machine learning models by supporting multi-layering, non-linear activation functions, and capable of capturing long-term dependencies. Deep neural networks provides excellent analytical and processing capacity to capture language semantics and syntax thus becoming closer to human sophistication.

Chalkidis and Kampas [12] in a survey discusses applications of deep learning for processing legal text-based of three different NLP tasks namely text classification, information extraction, and information retrieval. This work is primarily focusing on semantic feature representation for deep learning models. One of the important contributions of their research is the legal word embedding dataset using the word2vec model containing legislations from European countries. Bansal *et al.* [13] provides the comparative analysis of different legal tasks such as classification, summarization, case reviews, and predictions using deep learning models namely CNNs, RNNs, LSTM, and GRU. Their study is based on the classification of the legal task into three subdomains viz. data search, legal text analytics, and legal intelligent interfaces. They found that deep learning models provide state of the art performance for the majority of the studied systems.

Lippi *et al.* [14], [15] proposed a methodology to identify loopholes from online service agreements in the form of unfair clauses. They formulated the problem of identification of unfair clause a sentence classification problem with the experimental setup using support vector machines [16], combined with deep learning architecture i.e. convolution neural networks [17] and long-short term memory networks [18]. This work is available as a commercial tool for domain users.

Xia *et al.* [19] in their work emphasizes the need for intelligent justice through effective deep learning techniques. Considering the complex structure of legal documents, similarity analysis is a difficult task. To address this difficulty, they proposed an approach using the combination of Word2vec with legal document corpus to improve the accuracy of similarity analysis of law documents and demonstrated that their approach is showing improved performance.

Nanda *et al.* [20] extended their work based unsupervised lexical and semantic similarity techniques [21], [22] to evaluate multilingual legal corpus of European directives and national legislation (from Ireland, Luxembourg, and Italy). They used shallow neural networks to developed word and paragraph embedding models for the corpus. Proposed work develops unsupervised as well as supervised semantic similarity model to identify transpositions and their performance is evaluated on various feature sets.

Marques *et al.* [23] presented a scoring mechanism to rank the most relevant legal citation in case judgments to support the legal argument. The scoring mechanism developed for the system is using a feature matrix as each case article as a feature to classifier for recommendations. Another score value is making use of word embedding text similarity techniques for finding relevant citations. Researchers have claimed that their proposed technique is better in comparison to baseline techniques for ranking evaluation of relevance criteria.

3.3. Knowledge base population for legal text processing

Knowledge base is a machine-readable data repository in a structured format. Some of the popular commercially used knowledge bases projects include Wikidata [24], DBpedia [25], Freebase [26]. The graph is a well-suited data structure stores factual information in the form of relationships between entities. Knowledge base population (KBP) systems [27] extract knowledge from available resources and generate knowledge base by considering semantic and contextual information from the resources. Knowledge base population system's objective is to automatically identify entities from unstructured text documents and discovering the facts about those automatically extracted entities and represent it in a structured knowledge base format. A specific KBP system goal should be to use logical reasoning for drawing inferences based on the logical contents of the input data. KBP involves two separate sub-tasks, entity linking, and slot filling. The entity linking task [28] aligns textual mention of a named-entity to its appropriate entry in the knowledge base or determines that the entity does not exist in the KB. The slot filling task [29] collects information regarding certain attributes of an entity from the corpus. If the corpus does not provide any information for a given attribute, the system will generate a NIL response. Information Extraction is necessary and crucial for successfully populating knowledge bases.

The objective of IE from text corpus is to extract and represent information in a tuple of two entities and a relationship between them. The task of extracting information from a large no of documents in the absence of a Labeled data set is termed as open information extraction [30]. This paradigm is claimed to be portable across different domains. One can prefer open information extraction to analyze legal documents that run across several pages and can assists practitioners and ordinary people to get the essence of the complex legal document. One of the important challenges of the traditional information extraction approach [31] is the dependency on some handcrafted domain-specific pattern matching rules. Information extraction outside the boundary of pattern matching rules cannot be done using traditional IE. Refer to Table 1. Comparative analysis of traditional versus Open IE. Some of the error classes identified with IE [32] are the boundary errors class, uninformative extraction error class, redundant relations extraction error class, wrong extractions error class.

Table 1. Traditional Vs open IE

	Traditional IE	Open IE
Input	Text + Predefined relations	Text
Relation	Relations need to be defined in advance	Free discovery of relations
Extractor	Only predefined relations	All possible relation

A variety of approaches have been proposed to address entity linking and slot filling. These diverse approaches are providing new opportunities for both entity linking and slot filling tasks of KBP. TextRunner [33] and STANFORD OPENIE [34] is an example of OIE system. A knowledge graph is a very effective data structure for storing semantically related concepts together extracted by using open information extraction approach and represented using relational machine learning.

Shrinivasa *et al.* [35] developed a knowledge base named as crime base from online news articles in leading Indian newspapers Times of India and Deccan Chronicle from Jan 2018 to Jun 2018 as crime reports published in newspapers are more authenticate then info available on social media crime base contains crime entities from multiple modalities in machine-readable form which can be useful to law enforcement agencies for crime activities analysis and future predictions. The novelty of this work is considering the image as well as text data for the construction of a knowledge base. The crime base uses domain-specific manually crafted rule-based approach crime entities extraction by using techniques like Tokenization, POS tagging, NER, named entity disambiguation [36] contextual and semantic similarity measures [37] for text data, and low and high level for image data. The system visualization is don using OWL model [38]. Boella [39] proposed a legal knowledge management system for the understanding of legal terms, different norms, and interrelation between them. This system will be beneficial for legal experts as well as a common man for a better understanding of the legal domain. The main objective of the proposed work was to semi-automate the frequently needed tasks of classification of documents, get a clear understanding of legal terms, extracting key terms for the user query, and more sophisticated search options. The Semiautomatic knowledge population task in the legal domain [40] proposed is implemented using rule-based, statistical procedures for parsing from the Italian legal database of norms for sentence extractions followed by application of pattern matching rules to identify named entities from the corpus. Statistical framework and legislative XML are used to represent extracted named entities for visualization purpose.

3.4. Proposed approaches for Indian legal system

Legal information retrieval systems require to identify catchphrases from judgments automatically, a mechanism needs to be explored in depth. Mandal [41] proposed an approach using unsupervised learning, for extraction and ranking of catchphrases automatically using the noun phrases from judgments. The proposed system is compared with different supervised and unsupervised baseline systems and getting statistically better performance over those baseline systems. Like catchphrase detection, measuring similarity between different legal documents is also desired by IR systems, two types namely graph-based and text-based techniques are available for the said task. Mandal [42] proposed a similarity measuring approach for Indian legal documents using text-based methods combined with topic modelling and neural networks for word and document embedding for better results. This work proves that the embedding based approach outperforms over both graph-based and text based baseline systems. Bhattacharya [43] proposed an approach for automatic c identification of the rhetorical roles of sentences from Supreme Court of India judgments using deep neural networks. The significance of using deep neural network these systems work better than many baseline systems which use handcrafted features.

4. CONCLUSION

Legal text documents are structurally different than other domain texts such as news articles or bioinformatics domains. So, techniques adapted for information extraction from the legal domain demands for understanding the formats and semantics of the legal document. Also, legal documents exist in different varieties like contracts, reports, court judgments each of which follows a different layout and structure. From the survey conducted for IE from legal texts, it is very much visible that not much work is being carried out for IE for Indian law system documents. Though the NLP approach seems promising for legal text processing, representation of extracted information in machine-readable as well as user-friendly form creates a challenge for this approach. The deep learning approach is wildly adapted by the researcher community for various domains but creating a tagged corpus for this approach needs to much manual efforts for the complex and lengthy legal documents. After analysis of all the approached for legal text processing we find that the knowledge base population approach combined with NLP techniques can give promising results for legal text analysis for many tasks such as automatic summarization, finding the most relevant case judgments, classification of legal documents according to laws, acts or any other parameter, automatic head note generation for a case, finding the references for a given case through citations and may more for Indian law system. There is no. of areas open for exploration with different issues in legal text analysis where researchers from the Information Extraction community can contribute to benefit legal experts to get rid of the manual, complicated time-consuming task as well as a common man to better understand the legal domain. After discussing the need of automation in legal sector a fundamental question arises that whether automation in legal sector would replace the lawyer and legal analyst in future? To answer this question, one needs to understand that legal domain is highly driven by analysis, decision making, and representation techniques which is difficult to automate. Still there are some areas in legal domain where automation is highly desired. Due diligence-contract review, legal research conduction to save manual efforts. Prediction technology-to predict the probable outcome of the cases by analyzing previous judgments. Legal analytics-to generate the data points from past judgments, and identify relevant case laws to be used by lawyers in their present cases. Automation of documentation-by just submitting the relevant documents get your legal documents ready.

REFERENCES

[1] J. Ruhl, D. M. Katz, and M. J. Bommarito, "Harnessing legal complexity," *Science*, vol. 355, no. 6332, pp. 1377-1378, 2017, doi: 10.1126/science.aag3013.

[2] A. Kanapala, S. Pal, and R. Pamula, "Text summarization from legal documents: a survey," *Artificial Intelligence Review*, vol. 51, no. 3, pp. 371-402, 2019, doi: 10.1007/s10462-017-9566-2.

[3] T. Padayachy, B. Scholtz and J. Wesson, "An Information Extraction Model Using a Graph Database to Recommend the Most Applied Case," *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, 2018, pp. 89-94, doi: 10.1109/iCCECOME.2018.8658659.

[4] M. Surdeanu, R. Nallapati, and C. D. Manning, "Legal Claim Identification: Information Extraction with Hierarchically Labeled Data," *Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT-2010)*, 2010, pp. 1-8.

[5] M. G. Constantino *et al.*, "CLIEL: context-based information extraction from commercial law documents," *ICAIL '17: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pp. 79-87, 2017, doi: 10.1145/3086512.3086520.

[6] M. G. Buey, A. L. Garrido, C. Bobed, and S. Ilarri, "The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies," in *ICAART,* vol. 2, pp. 438-445, 2016, doi: 10.5220/0005757204380445.

[7] M. J. Bommarito, D. M. Katz, and E. M. Detterman, "LexNLP: Natural language processing and information extraction for legal and regulatory texts," *arXiv preprint arXiv:1806.03688*, pp. 1-9, 2018.
[8] J. Savelka, H. Xu, and K. D. Ashley, "Improving sentence retrieval from case law for statutory interpretation," *Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL*, pp. 113-122, 2019, doi: 10.1145/3322640.3326736.
[9] S. Kumar, P. K. Reddy, V. B. Reddy, and A. Singh, "Similarity analysis of legal judgments," *Compute 2011 - 4th Annual ACM Bangalore Conference,* vol. 17, pp. 1-4, 2011, doi: 17. 10.1145/1980422.1980439.
[10] I. Goodfellow, Y. Bengjo, A. Courville, and Y. Bengjo, "Deep learning," *MIT Press*, Cambridge 2016.
[11] Y. Goldberg, "Neural network methods in natural language processing," *Morgan and Claypool Publishers*, San Rafael, 2017.
[12] I. Chalkidis and D, Kampas, "Deep learning in law: early adaptation and legal word embeddings trained on large corpora," *Artificial Intelligence and Law*, vol. 27, pp. 171-198, 2019, doi: 10.1007/s10506-018-9238-9.
[13] N. Bansal, A. Sharma, and R. K. Singh, "A Review on the Application of Deep Learning in Legal Domain," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, vol. 559, pp. 374-381, 2019, doi: 10.1007/978-3-030-19823-7_31.
[14] Lippi M. *et al.,* "Automated detection of unfair clauses in online consumer contracts," in *Legal Knowledge and Information Systems: JURIX 2017: The Thirtieth Annual Conference*, Luxembourg, vol. 302, 2017, pp. 145-154.
[15] M. Lippi *et al.,* "CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service," *Artificial Intelligence and Law,* vol. 27, no. 2, pp. 117-139, 2019, doi: 10.1007/s10506-019-09243-2.
[16] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *ECML 98,* Springer, Berlin, Germany, 1998, pp. 137-142.
[17] Y. Kim, "Convolutional neural networks for sentence classification," *In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, Doha, Qatar, ACL, 2014, pp. 1746-1751, doi: 10.3115/v1/D14-1181.
[18] A. Graves and J. Scmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602-610, 2005, doi: 10.1016/j.neunet.2005.06.042.
[19] C. Xia, T. He, W. Li, Z. Qin and Z. Zou, "Similarity Analysis of Law Documents Based on Word2vec," *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2019, pp. 354-357, doi: 10.1109/QRS-C.2019.00072.
[20] R. Nanda, "Unsupervised and supervised text similarity systems for automated identification of national implementing measures of European directives," *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 199-225, 2019, doi: 10.1007/s10506-018-9236-y.
[21] R. Nanda, L. Di Caro, and G. Boella, "A text similarity approach for automated transposition detection of European Union directives," *In: 29th International conference on legal knowledge and information systems, JURIX 2016,* vol. 294, 2016, pp. 143-148, doi: 10.3233/978-1-61499-726-9-143.
[22] R. Nanda *et al.,* "A unifying similarity measure for automated identification of national implementations of European union directives," In *Proceedings of the 16th edition of the international conference on artificial intelligence and law*. ACM, 2017, pp 149-158, doi: 10.1145/3086512.3086527.
[23] M. R. S. Marques, T. Bianco, M. Roodnejad, T. Baduel and C. Berrou, "Machine learning for explaining and ranking the most influential matters of law," *ICAIL '19: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2019, pp. 239-243, doi: 10.1145/3322640.3326734.
[24] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledge base," *Communications of the ACM,* vol. 57, no. 10, pp. 78-85, 2014.
[25] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web,* Springer, Berlin, Heidelberg, 2007, vol. 4825, pp. 722-735, doi: 10.1007/978-3-540-76298-0_52.
[26] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247-1250, 2008, doi: 10.1145/1376616.1376746.
[27] Heng Ji and R. Grishman, "Knowledge Base Population: Successful Approaches and Challenges," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 1148-1158, 2011.
[28] X. Ling, S. Singh, and D. S. Weld, "Design Challenges for Entity Linking," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 315-328, 2015, doi: 10.1162/tacl_a_00141.
[29] M. Surdeanu *et al.*, "Stanford's Distantly-Supervised Slot-Filling System," *Theory and Applications of Categories,* 2011.
[30] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM,* vol. 51, pp. 68-74, 2008, doi: 10.1145/1409360.1409378.
[31] D.-T. Vo and E. Bagheri, "Open information extraction, Encyclopedia with semantic computing," *Semantic Computing*, pp. 3-8, 2016, doi: 10.1142/9789813227927_0001.
[32] R. Schneider, T. Oberhauser, T. Klatt, F. A. Gers, and A. Loser, "Analyzing Errors of Open Information Extraction Systems," *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, Association for Computational Linguistics, Copenhagen, Denmark, Sep. 2017.
[33] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open-domain information extraction," In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the*

7th International Joint Conference on Natural Language Processing, vol. 1, pp. 344–354, 2015, doi: 10.3115/v1/P15-1034.
[34] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A Review of Relational Machine Learning for Knowledge Graphs," in *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11-33, Jan. 2016, doi: 10.1109/JPROC.2015.2483592.
[35] K. Srinivasa and P. S. Thilagam, "Crime base: Towards building a knowledge base for crime entities and their relationships from online newspapers," *Information Processing & Management*, vol. 56, no. 6, 2019, Art. no. 102059, doi: 10.1016/j.ipm.2019.102059.
[36] G. Zhu and C. A. Iglesias, "Exploiting semantic similarity for named entity disambiguation in knowledge graphs," *Expert Systems with Applications*, vol. 101, pp. 8-24, 2018, doi: 10.1016/j.eswa.2018.02.011.
[37] R. Qu, Y. Fang, W. Bai, and Y. Jiang, "Computing semantic similarity based on novel models of semantic representation using Wikipedia," *Information Processing & Management*, vol. 54, no. 6, pp. 1002-1021, 2018, doi: 10.1016/j.ipm.2018.07.002.
[38] P. Buitelar, P. Cimiano, A. Frank, M. Hartung, and S. Racioppa, "Ontology-based information extraction and integration from heterogeneous data sources," *International Journal of Human-Computer Studies*, vol. 66, no. 11, pp. 759-788, 2008, doi: 10.1016/j.ijhcs.2008.07.007.
[39] G. Boella. L. Di Caro, and V. Leone, "Semi-automatic knowledge population in a legal document management system," *Artificial intelligence and Law*, vol. 27, pp. 227-251, 2019, doi: 10.1007/s10506-018-9239-8.
[40] E. de Maat, K. Krabben, and R. Winkels, "Machine learning versus knowledge-based classification of legal texts," in *Proceedings of legal knowledge and information systems conference: JURIX 2010.IOS Press*, 2010, pp. 87-96, doi: 10.3233/978-1-60750-681-2-87.
[41] A. Mandal, K. Gosh, A. Pal, and S. Ghosh, "Automatic catchphrase identification from legal court case documents," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2187-2190, doi: 10.1145/3132847.3133102.
[42] A. Mandal, R. Chaki, S. Saha, K. Ghosh, A. Pal, and S. Ghosh, "Measuring similarity among legal court case documents," in *Proceedings of the 10th annual ACM India compute conference*, 2017, pp. 1-9, doi: 10.1145/3140107.3140119.
[43] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, and A. Wyner, "Identification of rhetorical roles of sentences in Indian legal judgments," 2019, *arXiv preprint arXiv:1911.05405*.

BIOGRAPHIES OF AUTHORS

Ashwini V. Zadgaonkar is an assistant professor in the dept of CSE at RCOEM, Nagpur. She is graduated from B Tech Computer Technology and Mtech in CSE from Nagpur University and currently pursuing her Phd at RCOEM, RTMNU Nagpur, India. Her area of research is NLP and Data Mining.

Avinash J. Agrawal is an associate professor in the dept of CSE at RCOEM, Nagpur. He has done his BE from Nagpur university and Mtech in CSE from NIT Raipur He is a research Scholar of VNIT Nagpur. His research interests are the area of NLP, Data Mining and artificial intelligent. He has more than 50 publications in reputed journals and conferences.