NLP and AI Based Question Answering System in the Legal Domain

Abhinav Verma
Computer Science Department
Chandigarh University
Gharuan, Punjab
19bcs1232@gmail.com

Aman Saini
Computer Science Department
Chandigarh University
Gharuan, Punjab
19bcs1733@gmail.com

Er. Inakshi Garg
Computer Science Department
Chandigarh University
Gharuan, Punjab
inakshi.e12349@cumail.in

Abstract — The use of Question Answering (QA) systems in the legal domain has become increasingly popular as a means for people to seek legal advice. However, current QA systems often struggle to fully comprehend the legal context and provide answers that are relevant to specific jurisdictions due to the lack of domain expertise. This paper presents the development of a system for question answering specifically in the domain of law. The system utilizes legal data to automatically comprehend natural language queries from users and provides the most relevant answers to the given questions. Experimental results from a large-scale legal QA corpus demonstrate the effectiveness of the system. This system is expected to be highly important in closed domain systems and will be especially effective in the legal domain.

Keywords – Natural Language Processing (NLP), Artificial Intelligence (AI), Question Answering Systems (QAS)

I. INTRODUCTION

In recent years, the use of Question Answering Systems (QAS) has grown rapidly due to the increasing number of internet users. QAS is an advanced technology that helps users retrieve accurate answers to their natural language questions without the need for a query. QAS has made significant advancements in various languages such as English, Chinese, Japanese, Korean, etc. QAS is a specialized area within the field of information retrieval and there are many different types of QAS with their own unique application areas.

QA amalgamates various important fields like IR, NLP, Knowledge representation, logical reasoning, Machine learning and many more. [1] QAS can be used for many different purposes such as extracting information from documents, language learning, online examination systems, human and computer interaction, document management and classification of documents. QAS can also be classified into structured and semi-structured data depending on the source of answers. The primary objective of QAS is to retrieve answers to specific questions rather than full documents. QASs are information retrieval-based tasks that process questions posed in natural language by using pre-organized databases or a large corpus of documents published in natural language. [2]

There are two main types of QAS: open domain and closed domain. Open domain systems are web-based and have no restrictions on the era or domain while closed domain systems have limited work domains such as medicine or weather forecasting. [3] QAS has become an essential tool in many different areas and has the potential to solve complex problems and improve efficiency in various industries.

Closed Domain Systems :

A closed domain system refers to a system that deals with a particular domain or topic. This type of system makes it easier to ask questions and retrieve answers since NLP (Natural Language Processing) systems are adept at finding specific topic questions and retrieving the answers. Closed domain systems can work very quickly when the questions are limited to a specific domain or topic. Descriptive questions are more effective than procedural ones for this type of system. Machine reading applications have also been created in the medical field such as for Alzheimer's disease.

Open Domain Systems :

In information retrieval, an open domain QA system aims to provide accurate answers to user-generated questions. The system returns short, fragmented text instead of documents. Information retrieval, computational linguistics, and knowledge representation are utilized to find the answers to the given questions. The system takes natural language questions as input rather than a group of keywords, making it more user-friendly. However, it is more difficult to implement since there are many types of questions and it is challenging for the system to identify the correct question type to provide the appropriate answer. Assigning each question type to a question is a critical and challenging task. The success of the retrieval process depends on identifying the correct question type and providing the correct answer type to yield an accurate response.

As the name suggests, question answering can be open-ended and pertain to any domain or topic. Therefore, the system must have a large database of information, allowing answers to be retrieved at any time.

Documents such as legal depositions comprise conversations between a set of two or more people with the goal of identifying observations and the facts of a case. [4]. Question answering systems can be a valuable tool in the legal domain by assisting lawyers, judges and other legal professionals in accessing and analyzing large volumes of legal information quickly and accurately. Here are some examples of how question answering systems can be used in the legal domain such as Legal Research where a question answering system can be used to search and retrieve relevant case laws, statutes, regulations and other legal documents from large legal databases. By providing a natural language query, the system can quickly narrow down the search results and present the most relevant information to the user. In document analysis where a question answering system can be used to analyze legal documents such as contracts, pleadings and briefs. The system can extract key information such as parties involved, relevant dates and legal concepts and answer questions related to the content of the document. In legal advice where a question answering system can assist in providing legal advice to clients by answering questions related to their legal issues. For example, a client might ask a question about the potential consequences of a certain action or the requirements for filing a particular type of legal claim. In case preparation where a question answering system can assist lawyers in preparing for a case by answering questions related to the legal issues involved. The system can help identify relevant legal precedents and provide insights into the interpretation and application of the law. A legal QA system must also handle questions where no single answer exists. [5] Overall, question answering systems can save time and increase the efficiency of legal research and analysis, making it easier for legal professionals to access and understand the vast amount of legal information available.

II. LITERATURE SURVEY

Question answering is an area of computer science that combines natural language processing and information retrieval. It involves building systems that can understand and respond to questions posed by humans in natural language format. Typically, a computer program will generate its own answers by querying a structured database of information, also known as a knowledge base or by extracting relevant information from unstructured datasets such as reference texts, internal documents, webpages, newswire reports, Wikipedia pages or subsets of the World Wide Web. A QA system enables users to access the knowledge resources in a natural way (i.e., by asking questions) and to get back a relevant and proper response in concise words. [6]

The first QA systems were developed to access data over databases in the late sixties and early seventies. [7] In the early stages of question answering, two systems called BASEBALL and LUNAR were developed. BASEBALL was able to answer more questions over a one-year period while LUNAR was designed to provide geological analysis of rocks returned from Apollo moon missions. These systems were highly effective and demonstrated the potential of question answering. In subsequent years, other systems such as ELIZA, DOCTOR and SHRDLU were developed, each with its own unique focus and capabilities. One of the earliest QA systems was ELIZA, developed in 1964. [8]

In the 1970s, knowledge bases were developed to organize and streamline domains of knowledge. These expert systems served as interfaces for question answering systems which were designed to extract information from them. This led to the development of computational linguistics in the 1970s and 1980s, which has since enabled more advanced and effective question answering systems. For instance, the EAGLI system has been developed for the health and life benefits sector. [9]

Question answering systems heavily rely on high-quality search corpora and larger collections of data typically yield better performance. Fueled by the wide scale expansion of digitized legal texts, the prospect for deploying cutting-edge NLP techniques within the legal domain has become increasingly possible. [10] In order to optimize data redundancy, it is important to ensure that small pieces of information are presented in multiple ways, allowing for easier retrieval. When all relevant information is available in a variety of formats, it becomes easier for the QA system to tackle complex tasks. Additionally, having accurate data located in close proximity to incorrect data can help to identify and rectify errors. Ultimately, the effectiveness of a QA system is largely dependent on its reasoning power. Many QA systems have been developed using Prolog, a logic programming language closely associated with artificial intelligence.

QA systems are now determined in search engines like google and phone conversational interfaces. [11] Question answering systems have been developed in the last few decades and some of them are

SHRDLU was a natural language processing program that could manipulate blocks in a virtual world. It was able to understand natural language commands and could answer questions about the virtual world.

MYCIN was a rule-based expert system that provided advice on the diagnosis and treatment of bacterial infections. It was one of the first successful expert systems and was able to answer questions based on its knowledge of the medical domain.

IBM Watson is a question answering system that was developed by IBM to compete on the quiz show Jeopardy!. It uses natural language processing, machine learning and other techniques to answer questions posed in natural language.

Siri is a virtual assistant developed by Apple that uses natural language processing to answer questions, perform tasks and provide recommendations.

Google Assistant is another virtual assistant that uses natural language processing to answer questions and perform tasks. It is integrated into Google's search and other products and can be accessed through smartphones, smart speakers and other devices.

However, these returned lengthy documents and cannot provide an exact solution to the user's problem and it may be time consuming to review all of them, without having a guarantee of finding the desired answers. [12]

III. PROPOSED METHODOLOGY

A simple question answering system is created in Python using natural language processing (NLP) techniques. Here's a high-level overview of how it works:

Data Collection:

Collect a set of text documents that contain the information you want the system to answer questions about. This can be done through web scraping or using pre-existing datasets.

Data collection is a critical step in developing a question answering system (QAS). It involves gathering a large amount of data from various sources such as web pages, databases, and other documents, that can be used to train and test the QAS.

The data collected for a QAS should include a diverse range of questions and corresponding answers. The questions should be phrased in natural language and cover different topics, while the answers should be accurate, relevant and informative.

Data collection for a QAS can be done in various ways, such as web scraping, manual curation or using existing datasets. Web scraping involves automatically extracting data from web pages using software tools. Manual curation, on the other hand, involves manually selecting and annotating data from various sources. Existing datasets, such as the Stanford Question Answering Dataset (SQuAD), can also be used to train and evaluate a QAS.

Once the data is collected, it needs to be preprocessed and cleaned to remove irrelevant information and ensure consistency. This involves techniques such as text normalization, entity recognition and entity linking.

Overall, data collection is a crucial step in developing a QAS, as the quality and diversity of the data collected will have a significant impact on the performance of the system.

Data Preprocessing:

Preprocess the text data by tokenizing, stemming and removing stop words. This is done to ensure that the system can understand the text data.

Data preprocessing is an essential step in the development of a question answering system. It involves preparing the data in a way that makes it usable for the system to understand and process.

In the context of a question answering system, data preprocessing typically involves several steps:

Corpus Cleaning: The first step is to clean the corpus of any irrelevant or redundant data. This can involve removing special characters, punctuation, stop words and any other data that is not necessary for the question answering task.

Tokenization: The next step is to tokenize the text data, which involves breaking it down into smaller units such as words, phrases or sentences. This process helps the system understand the structure of the text and enables it to identify the relevant parts of the text that are related to the question.

Lemmatization and Stemming: Lemmatization and stemming are techniques used to reduce the number of different word forms in the text data. This helps to normalize the text and make it easier for the system to match the text with the question.

Named Entity Recognition: Named entity recognition involves identifying and classifying named entities in the text, such as people, organizations, and locations. This can be useful in answering questions that require specific information about these entities.

Sentence Boundary Detection: The final step is to detect sentence boundaries, which involves identifying where one sentence ends and another begins. This is important for answering questions that require information from multiple sentences.

By performing these preprocessing steps, the question answering system can extract the relevant information from the text and provide accurate answers to questions.

Question Processing:

Process the user's question by tokenizing, stemming, and removing stop words. Each question is converted to lowercase, punctuation symbols are eliminated (parenthesis, hyphens, numbers, etc.). [13] This is done to ensure that the system can understand the question.

Question processing is a crucial step in building a question answering system. It refers to the process of analyzing and understanding a natural language question and transforming it into a structured form that can be used to retrieve relevant information from a knowledge source or a database. The Question Processing results are a list of keywords plus the information for the asking point. [14]

The goal of question processing is to identify the type of question being asked, extract the relevant keywords and phrases, and determine the appropriate query language to retrieve the answer. This involves several sub-tasks such as parsing the question into its constituent parts, identifying named entities, determining the semantic relationships between words and disambiguating word senses.

Once the question has been processed, the question answering system can use this information to formulate a query and retrieve relevant information from its knowledge source or database. The retrieved information is then analyzed and synthesized to generate a concise and accurate answer to the original question.

Overall, question processing is a critical component of a question answering system, as it directly impacts the accuracy and completeness of the answers provided.

Information Retrieval:

Retrieve relevant information from the text documents based on the user's question. This can be done using techniques such as keyword matching, named entity recognition or information retrieval algorithms.

Information retrieval (IR) is a key component of question answering (QA) systems. It refers to the process of retrieving relevant documents or pieces of information from a collection of data sources in response to a user's query or question. Finding information in a large text database is the traditional subject of study in information retrieval. [15]

In a QA system, the user's question is typically treated as the query, and the system searches for the most relevant answer(s) within a collection of documents or knowledge sources, such as a database or the internet. The IR component of the system is responsible for selecting and ranking relevant documents based on their content, similarity to the query and other factors such as the authority or trustworthiness of the source.

IR techniques used in QA systems may include natural language processing (NLP), indexing, document ranking and query optimization. These techniques aim to retrieve the most accurate and relevant information for the user's query, with the goal of providing a concise and accurate answer to the question.

Answer Extraction:

Extract the answer from the retrieved information. This can be done using techniques such as pattern matching, syntactic parsing or machine learning algorithms.

Answer extraction is a key component of a question answering (QA) system that involves identifying the most relevant information in a given text or set of texts to answer a user's question. In other words, it is the process of identifying and extracting a short and concise answer to a user's question from a larger pool of information, for example responding 29,029 feet to a question like "How tall is Mt. Everest?". [16]

Answer extraction algorithms typically use a combination of natural language processing (NLP) techniques, such as part-of-speech tagging, named entity recognition and semantic parsing, to identify and extract the most relevant information from the input text. The extracted answer is then presented to the user as a direct response to their question.

Answer extraction can be challenging, as the relevant information may be spread across multiple sentences or paragraphs or may be expressed in a way that is not immediately obvious. Additionally, the answer may need to be synthesized from multiple pieces of information or may require inference or logical reasoning to derive. However, with the advancements in deep learning and NLP technologies, answer extraction systems have become increasingly accurate and effective.

Answer Generation:

Generate a final answer to the user's question based on the extracted answer. This can involve formatting the answer in a readable way or providing additional context.

Answer generation is the process of creating a coherent and relevant response to a user's question in a question answering system. It involves not only extracting the relevant information from a given text, but also generating a grammatically correct and informative answer in natural language that addresses the user's query. In QA, answer formulation does: improving AE or improving human-computer interaction (HCI). [17]

Answer generation can involve different approaches depending on the type of question answering system. For example, in a fact-based QA system, the answer may be generated by simply selecting the most relevant fact or piece of information from the input text, while in a more complex system, such as a conversational QA system, the answer may involve generating a longer and more detailed response based on a combination of extracted information, external knowledge sources and context.

Answer generation in question answering systems can be accomplished using a variety of techniques, such as rule-based systems, template-based systems, and more advanced machine learning models, such as neural networks and deep learning models. These systems can be trained on large amounts of data to learn patterns in the text and generate more accurate and relevant responses to user queries.

Overall, answer generation is a critical component of any question answering system, as it is what allows the system to provide users with actionable information in a form that is easy to understand and use.

IV. FUTURE SCOPE

The future scope of question answering systems in the legal domain is promising. As the amount of legal information continues to grow rapidly, it becomes increasingly challenging for legal professionals to keep up with the latest developments and to search and retrieve relevant information efficiently. Question answering systems that use natural language processing (NLP) and machine learning techniques have the potential to improve legal research and make it more efficient. They can help lawyers and legal professionals to quickly find the relevant legal cases, statutes and regulations that they need for their work. In addition, question answering systems can also help to automate some of the more routine legal tasks, such as document review and contract analysis. This can save time and reduce costs for law firms and their clients. Another potential application of question answering systems in the legal domain is in the area of legal chatbots. These could be used to provide basic legal information and advice to individuals who cannot afford to hire a lawyer or who need immediate assistance outside of normal business hours. Overall, the future scope of question answering systems in the legal domain is promising and we can expect to see more advanced and sophisticated systems being developed in the coming years.

CONCLUSION

The QA system can serve as a learning companion and be used in the field of law to solve closed domain problems and provide relevant answers to questions. With the need for self-paced learning in the global era, the system can construct systems that can evaluate or grade answers with results consistent with human performance. However, there are several challenges in implementing this system, including knowledge representation, precise representation for proper understanding, paraphrasing, conceptual learning, online accessing of descriptive questions, evaluation of answers, inclusion of figures, tables and mathematical equations. These challenges are complex and finding solutions for them is difficult, but the application needs are high. Therefore, there is great potential for exploring the challenges in the QA field to improve the system that can help in the legal domain to a great extent.

ACKNOWLEDGEMENT

We express our heartfelt gratitude to Professor Inakshi Garg from the Computer Science Department at Chandigarh University, Gharuan for providing us with her valuable guidance and motivation throughout the course of this work.

REFERENCES

[1] Bhushan Zope, Sashikala Mishra, “Question Answer System: A State-of-Art Representation of Quantitative and Qualitative Analysis”, Big Data Cogn. Comput. 2022, 6, 109

[2] Question Answering Systems: A Systematic Literature Review, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 3, 2021

[3] Ajitkumar M. Pundge, Khillare S.A., C. Namrata Mahender, “Question Answering System, Approaches and Techniques: A Review”, International Journal of Computer Applications (0975 – 8887) Volume 141 – No.3, May 2016

[4] Saurabh CHAKRAVARTY, “Improving the Processing of Question Answer Based Legal Documents”, Virginia Tech, Blacksburg, VA 24061

[5] A Free Format Legal Question Answering System, Natural Legal Language Processing Workshop 2021, pages 107–113 November 10, 2021

[6] Sanjay K Dwivedi, Vaishali Singh, “Research and reviews in question answering system”, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, Procedia Technology 10 (2013) 417 – 424

[7] Dennis Diefenbach, “Core Techniques of Question Answering Systems over Knowledge Bases: a Survey 2017”, Université de Lyon, Published in Knowledge and Information Systems.

[8] The Question Answering System Using NLP and AI. International Journal of Scientific & Engineering Research Volume 7, Issue 12, December-2016 ISSN 2229-5518

[9] A. Clementeena, Dr. P. Sripriya, “A literature survey on question answering system in natural language processing”, International Journal of Engineering & Technology, 2018

[10] Daniel Martin Katz, Dirk Hartung, “Natural Language Processing in the Legal Domain”

[11] Question Answering System Using Natural Language Processing, International Journal of Research in Engineering, Science and Management, VOL. 4, NO. 12, DECEMBER 2021

[12] Weiyi Huang, Jiahao Jiang, “AILA: A Question Answering System in the Legal Domain”, Proceedings of the Twenty-Ninth

International Joint Conference on Artificial Intelligence (IJCAI-20) Demonstrations Track

[13] Alfredo Monroy, “NLP for Shallow Question Answering of Legal Documents Using Graphs”, Conference Paper · March 2009

[14] Rohini Srihari, Wei Li, “A Question Answering System Supported by Information Extraction”, supported in part by the SBIR grants F30602-98-C-0043 and F30602-99-C-0102 from Air Force Research Laboratory (AFRL)/IFED.

[15] Hwee Tou Ng, “Question Answering Using a Large Text Database: A Machine Learning Approach”, DSO National Laboratories.

[16] Daniel Jurafsky & James H. Martin, “Chapter 25 - Question Answering, Speech and Language Processing”

[17] Bolanle Ojokoh, Emmanuel Adebisi, “A Review of Question Answering Systems”, Journal of Web Engineering, Vol. 17 8, 717–758.