Intelligent Q&A System For Public Legal Services

Gezhong Pan
College of Mathematics
Sichuan University
Chengdu, China
1405086583@qq.com

Yihan Gu
Law School Sichuan
University Chengdu,
China
2911169633@qq.com

Wujie Xiong
College of Computer
Science Sichuan
University Chengdu,
China
270516357@qq.com

Weisheng Zhong*
Union Big Data Technology Co.,Ltd
Chengdu, China
zhongweisheng@unionbigdata.com

***Abstract*—With the development of the economy and society, the change in social results and life order in less developed areas, the civil disputes begin to show the trend of diversification and complexity. At the same time, due to the unbalanced and inadequate development in the less developed areas, the supply of legal services in these areas is difficult to meet the growing demand for legal services.**

The advent of the Internet era has brought new vitality to the rule of law. The appearance of the dialogue system pro- vides the foundation for realizing intelligent question-answering technology. Intelligent question-answering technology mainly in- cludes two important parts: information retrieval and answer extraction. The information retrieval of legal intelligent question answering system considers the combination of common question-answering database(FAQ) and network retrieval. The answer extraction algorithm uses deep learning model text as semantic representation and is applied to FAQ. The realization of these technologies provides an equal, convenient and accurate way for the popularization and publicity of public legal services.

***Index Terms*—Public Legal Services, Intelligent Q&A, Artifi- cial Intelligence**

I. INTRODUCTION

Public legal service is an important part of government public service function. To implement the construction of the public legal service system, build a modern public legal service system, and solve the problem of the large demand for legal services in less developed areas but a shortage of resources[1, 2] , the idea of using intelligent Q&A system to realize online legal aid has been recognized by the society as soon as it appeared. And the basis of intelligent Q&A system is the dialogue system[3] .

Nowadays, people pay more and more attention to the dialogue system in various fields. Specifically, the dialogue system can be roughly divided into two types : task-oriented dialogue system and non-task-oriented dialogue system (also known as chatbot[4]). Among them, the task-oriented system which I depend on aims to help users complete actual specific tasks, such as helping users find goods, booking hotels and restaurants. One approach to the widespread application of task-oriented systems is to treat the dialogue response as a pipeline. The system first understands the information con- veyed by human as an internal state, then takes a series of cor- responding behaviors according to the strategy of the dialogue state, and finally converts the actions into the expression form of natural language. While language understanding is handled through statistical models, most deployed conversational sys- tems still use manual features or manual rules for state and action space representation, intent detection, and slot filling.

The task-oriented dialogue system that we refer to is the open-source Convlab-2 project of Tsinghua University profes- sor Minglie Huang[5] .The open-source can be downloaded from https://github.com/thu-coai/ConvLab-2. Convlab-2 is an open- source toolkit that enables researchers to build task-oriented dialogue systems with state-of-the-art models, perform an end- to-end evaluation, and diagnose the weakness of systems. Through ConvLab-2 technology, we first combined artificial intelligence and legal analysis, using advanced machine learn- ing algorithms based on natural language processing (NLP) to respond to user needs for legal resources. Through BERT deep learning model, we can mine the effective semantic in- formation conveyed by users and identify the case knowledge elements, to solve the problem of semantic understanding of legal data[6] , and then give rapid and correct feedback.The two important algorithms support the intelligent question- answering system of public legal service.

To develop this intelligent question-answering technology for public legal services, we mainly follow three steps: data preprocessing and training, intelligent question-answering sys- tem, and scene testing. After the completion of the project, the terminal software was applied by the Unionbigdata Company and then the user was asked about the usage and received feedback, and the product was modified to improve the effect in practical application. Now we plan to promote the use of this product after improvement.

II. DATA PREPROCESSING AND TRAINING

A. Data Sources

First of all, the database includes several open-source legal literature databases such as the National Laws and Regu- lations Database (https://flk.npc.gov.cn/), China Judgements Online (https://wenshu.court.gov.cn/) and so on. In addition, with authorization, the project obtained data from Chengdu Unionbigdata Company and China Judicial Big Data Service Platform, which provided extensive original data for sample set establishment and model training, ensuring the universality of the model.

B. Data Cleaning

Legal data preprocessing aims to automatically structuralize semi-structured data based on the requirements of specifi- cations and to transform different forms of legal data into more recognizable and standard text data. Since the data in this project are legal provisions and mostly dialogues between people, the sample features are obtained manually, and then the clustering segmentation method is used to clean the data with the machine learning algorithm.

C. Text And Training

Text representation is to effectively mine the implicit knowl- edge in multi-source heterogeneous data of legal documents to support the data application of the platform. To support the application of legal data in this project, we focuse on the deep semantic learning method for legal data. Through BERT deep learning model, semantic information in texts is mined and case knowledge elements are identified, to effectively solve the problem of semantic understanding of legal data.

General text representation is divided into discrete represen- tation and distributed representation. The typical representative of discrete representation is the word bag model, which repre- sents documents as a collection of feature items (words), and solves the problem that it is difficult for classifiers to process discrete data. However, this approach ignores the relation of words in context, and there are problems such as sparse data, high vector dimension, and the relationship between words that cannot be measured. Therefore, we adopt the BERT model of text distributed representation.

The BERT model significantly improves the performance of natural language processing tasks including text classification, by jointly adjusting the context and pre-training the deep bidi- rectional representation in unlabeled text. BERT is a typical self-coding Model, and its pre-training process adopts the self-coding idea of noise reduction, that is, the MLM (Masked Language Model) mechanism, which randomly blocks some words for pre-training of word vector. At the same time, the combination of token vector, segment vector, and posi- tion vector is introduced in the token representation of each position, which can carry out more comprehensive semantic representations. In the meanwhile, through the mechanism of mask, the learning process of word vector can introduce the context information at the same time in a single training, instead of the rigid stitching way of bidirectional RNN, and its feature extraction ability is far greater than the model of RNN and CNN. The greatest contribution of the BERT model is that it can obtain bidirectional context information[7] .

At present, natural language processing technology basically needs to build a network model to complete a specific task. First, you need to randomly initialize the parameters, and then start training the network, tweaking it until it loses less and less. The initial initialization parameters change throughout training. When you are satisfied with the results, you can save the parameters of the training model so that the trained model can achieve better results the next time you perform similar tasks. This process is pre-training, and BERT is the most popular text pre-training model. Later, you receive a similar text task. At this point, you can directly use the previously saved model parameters as the initialization parameters of the task, and then modify the results as you go through the training. You’re using a pre-trained model this time, and the process is fine-tuning.

D. Text Categorization

Shallow learning models usually need to obtain good sample features by manual methods and then classify them using classical machine learning algorithms. Therefore, the effec- tiveness of this method is largely limited by feature selection and feature extraction. However, unlike shallow models, deep learning integrates feature engineering into the model fitting process by learning a set of nonlinear transformations used to map features directly to the output.

The deep learning model avoids the artificial design of rules and features and automatically provides semantically meaningful representation for text mining[8] . Therefore, most research work on text classification is based on a deep neural network, which is a data-driven method with high computa- tional complexity. In this process, the BERT model is used to extract semantic feature information. The advantage of the BERT model is its strong ability to extract features, which solves the difficulty that general word vector methods cannot deal with polysemy.

III. INTELLIGENT Q&A SYSTEM

A. Information Retrieval

The information retrieval of the intelligent question-answering system based on public legal service considers the combination of FAQ and network retrieval. FAQ database retrieval refers to the calculation of the similarity between the questions entered by the user and the questions in the FAQ database. If the maximum human and similarity are higher than a threshold set manually, the retrieval is successful and the answer to the maximum similarity question is returned to the user. Web search refers to the use of search engines to search for information related to the answer according to the keywords of the question and its extension. In the process of information retrieval, the question-answering system retrieves the FAQ library first and then retrieves the network when the FAQ library fails. However, considering that there are many errors in network retrieval, a paragraph keyword relation database (PKR) is established for retrieval[9] .

B. Answer Extraction

In practical problems, similar structures are used to express sentences with the help of questions and answers. However, there is a certain interaction between the answer and the question representation, that is, the information of the question is combined in the process of obtaining the representation of the candidates answer. In the representation of candidate answers, the attention mechanism gives more weight to the words related to the question. Through experimental compar- ison, this project adopts deep learning model text to make semantic representation and applies it to FAQ. Compared with the traditional model, the Q&A system and the final answer extraction module have the advantages of saving a lot of manpower and material resources for manually extracting features, because they automatically extract the relationship between words from a large number of samples, and they can combine the structural information in phrase matching with the hierarchical nature of text matching to discover features that are difficult to discover in traditional models.

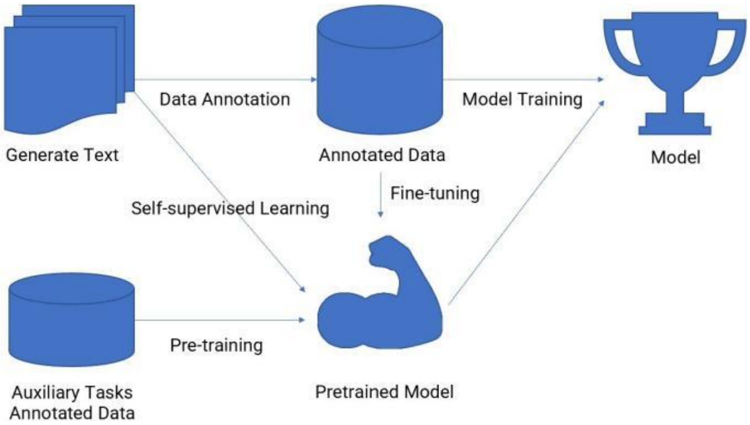

Fig. 1. Text preprocessing and training process

IV. SCENE TESTING

A. Efficiency Test

The program was trial-run in a laboratory environment to test the high efficiency of data. At the same time, the dialogue data provided by Chengdu Unionbigdata Company is used for high-pressure tests of the platform, and the completion time is used to determine the efficiency of the program. This reflects the superiority and universality of the program.

B. Robust Test

Based on the previous data and hardware, the function and system of each module are tested for fault tolerance, and error data processing, abnormal condition processing, and illegal operation processing are designed to detect whether the software module can maintain normal work under abnormal input and harsh environmental conditions. Ensure that each functional module of the terminal software works properly when processing error data and abnormal problems, and improve the fault tolerance of the terminal software[10] .

(1) Error Data Processing: The test method of error data processing is to manually input illegal data to the specified module and check whether the response and prompt informa- tion of terminal software are normal. The specific test methods are as follows:

(a) Determine which values are illegal according to the test specifications;

(b)Manually input the corresponding illegal data for testing;

(c)Check the test results. Usually, the test results will automatically adjust the data and give the correct prompt information;

(2) Abnormal Condition Processing: To test and handle exceptions caused by non-human factors and check whether the test terminal can handle exceptions properly. Exceptions include network exceptions, server exceptions, and terminal software exceptions.

Network Exception

Network exceptions include network congestion or interfer- ence, 4G/5G network switchover, circuit domain/data domain service conflict, and communication network unavailable.The specific test methods are as follows:

(a) Use the shielding box to simulate the situation of disconnection or weak signal, or set the situation that the network of terminal software is inconsistent with the actual network;

(b) Test the corresponding network-related functions.

For example, if the terminal software network is set to TD only, perform the test in an area without TD signals. The terminal software is required not to generate garbage data when a network exception occurs and to resume previously interrupted operations when the network is available.

Server Exception

Simulate server exceptions by using a test platform. When some test items cannot be tested in the live network environ- ment, change to the laboratory card and laboratory network environment for testing. The specific test methods are as follows:

(a) Modify data on the test platform;

(b) Test relevant data.

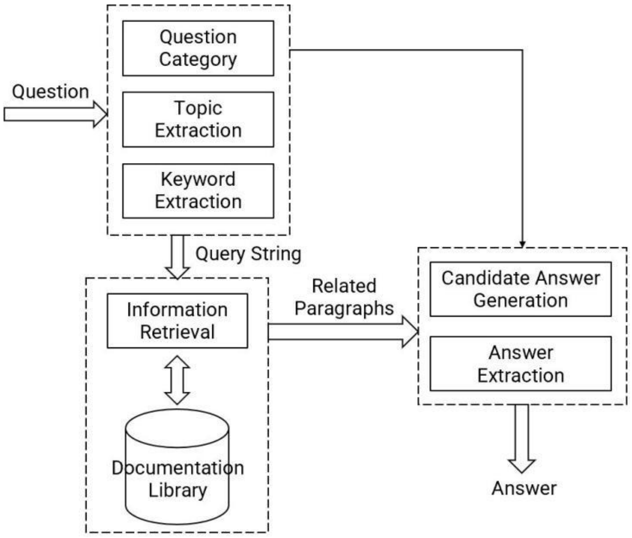

Fig. 2. Answer extraction algorithm

Terminal Software Exception

Terminal software abnormality mainly tests the case of low power. Use the battery of about 15 % to enter each module for testing, and check the working condition of the module in the case of low power. A low power warning should not harm the ongoing operation. The specific test methods are as follows:

(1) Test relevant modules when the battery level is about 15%;

(2) Test relevant modules when the battery level is about There will be a low power reminder when the power is lower than 10%, which will not interfere with the ongoing operation.

(3) When the power is lower than 0%, the terminal software shuts down automatically. Check that the shutdown will not lead to data loss, and then check whether the charging can continue.

C. Illegal Operation Processing

Illegal operation processing refers to the abnormal oper- ation of terminal software caused by human factors while operating the basic functions of the module. These illegal operations include modifying and deleting data during data synchronization, modifying and deleting data during large- volume data transmission, and upgrading terminal software when disturbed by external factors. Unauthorized operations on terminal software should not generate garbage data, and edited data can be saved to ensure no data loss.The specific test methods are as follows:

(a) Establish the necessary test environment or preset con- ditions;

(b) Perform illegal operations;

(c) Check the test results which should not generate garbage data and save the data which has been edited to ensure that no data loss.

V. CONCLUSION

Our team aims to build an intelligent Q&A system for public legal services in economically underdeveloped areas, and on this basis further establish a full-service legal platform integrating the legal service demand-side and the legal service supply-side. The legal service demand-side includes AI legal Q&A and grassroots legal services, and the legal service supply-side provides a platform for legal service providers to provide public legal services such as legal consultation, legal aid, and people’s mediation.

Our original intention is to promote the sinking of legal service resources to vulnerable groups and promote the match- ing of legal service demand and supply. In different modules, the law, artificial intelligence, and big data technology are combined to conform to the upsurge and trend of Internet court trials, so that the people in less developed areas can rely on the law to solve conflicts and disputes and improve the people’s sense of access to legal service resources.

REFERENCES

[1] Huang Dongdong & Zhang Rui. (2021). Digital Technology, National Governance, and Public Legal Service System reform. Study BBS (03): 121-130. Doi: 10.16133 / j.carol Carroll nki XXLT. 2021.03.017..

[2] Huang Dongdong, & Zhang Na. (2020). Research on equalization of legal aid from the perspective of basic Public Legal Services. Shandong Social Sciences (6), 5.

[3] Ma D N. Research on the development of intelligent law from the perspective of literature review.

[4] Chen, H. , Liu, X. , Yin, D. , & Tang, J. . (2017). A survey on dialogue systems: recent advances and new frontiers. arXiv e-prints.

[5] Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., ... & Huang, M. (2020). Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. arXiv preprint arXiv:2002.04793.

[6] Wen, T. H. , Vandyke, D. , Mrksic, N. , Gasic, M. , Rojas-Barahona, L.

[7] Nakano, M. , & Komatani, K. . (2020). A framework for building closed- domain chat dialogue systems. Knowledge-Based Systems, 204, 106212.

[8] Zhao, Y. J. , Li, Y. L. , & Lin, M. . (2019). A review of the research on dialogue management of task-oriented systems. Journal of Physics Conference Series, 1267, 012025.

[9] Devlin, J. , Chang, M. W. , Lee, K. , & Toutanova, K. . (2018). Bert: pre- training of deep bidirectional transformers for language understanding.

[10] Yuan, C. , Xue, M. , Zhang, L. , & Wu, H. . (2019). Robustness Analysis on Natural Language Processing Based AI Q&A Robots.