Research on Chinese Legal Document Reading Comprehension Based on Pre-Training Model

Lufeng Yuan
Beijing China-Power Information, Technology Co. Ltd, Beijing China
yuan_lufeng@163.com

Wei Zhang
Beijing China-Power Information, Technology Co. Ltd, Beijing China
853160858@qq.com

Xiaoxin Gao
Beijing China-Power Information, Technology Co. Ltd, Beijing China
13521395850@163.com

Linlin Zhao
Beijing China-Power Information, Technology Co. Ltd, Beijing China
zhaolinlin@sgitg.sgcc.com.cn

Bin Liu
Beijing China-Power Information, Technology Co. Ltd, Beijing China
levense_lb@163.com

Maokai Liu
Beijing China-Power Information, Technology Co. Ltd, Beijing China
milkai6@aliyun.com

ABSTRACT

We research how to read and understand Chinese legal documents. At first, we analyze the difficulties of reading comprehension of Chinese legal documents. Data imbalance exists seriously among span extraction query, yes/no query and unanswerable query, that is span extraction queries account for more than 80%. The reading comprehension of Chinese legal documents is a typical long text reading problem. Then we propose a framework for reading and understanding Chinese legal documents. Based on the Bert pre-training model, the framework performs fine-tine for Chinese legal documents, adopts a variety of deep learning models, and uses data enhancement and ensemble strategy to solve reading comprehension of Chinese legal documents. Finally, we test the framework with real legal documents, and the macro average F value can reach 82.773.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**; • **Information extraction**;

KEYWORDS

Chinese legal document, Reading comprehension, Bert, Pre-training model, Data imbalance, Ensemble strategy

ACM Reference Format:
Lufeng Yuan, Wei Zhang, Xiaoxin Gao, Linlin Zhao, Bin Liu, and Maokai Liu. 2021. Research on Chinese Legal Document Reading Comprehension Based on Pre-Training Model. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487157

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487157

1 INTRODUCTION

Reading comprehension is an important frontier topic in the field of natural language processing and artificial intelligence. Legal documents contain a lot of case information, such as time, place, character relationship and so on. By automatically reading and understanding of legal documents, judges, lawyers and the public can obtain the required information more quickly and conveniently. Therefore, it is of great significance to apply reading comprehension technology to legal document analysis.

Hermann et al [1]. proposed the data set CNN/Daily Mail in 2015 for the task of machine reading comprehension, which contains more than 1.26 million data. Based on CNN/Daily Mail, Stanford University has improved the Attentive Reader [2]. By using bilinear as the attention and analyzing only entities, the performance of a single neural network model is improved by 5%.

In order to solve the limitation that CNN/Daily Mail is a cloze data set, Rajpurkar et al. proposed SQuAD [3] (Stanford Question Answering Dataset) in 2016. The data set contains 107785 triples with question, context, answer, and context comes from 536 Wikipedia articles. Since the emergence of SQuAD, a large number of representative models have appeared. Caiming Xiong et al. introduced the Dynamic Coattention Network (DCN) [4] for question answering which won the ninth place in the SQuAD competition. A single DCN model increased the previous F1 value from 71.0% to 75.9%, and increased to 80.4% after using model ensemble. In 2017, Microsoft [5] adopted R-Net to model reading comprehension task and won the first place in SQuAD competition. In 2018, Alibaba's SLQA(Semantic Learning for Question Answering) [6] achieved the first place in the SQuAD competition. This is the first time that the accurate matching score of machine reading comprehension exceeds the result of human beings. Subsequently, the V-Net model [7] [8] proposed by Baidu reached the peak of Microsoft Marco dataset. Baidu won the first place only by single model, and did not submit the results of multi model ensemble. Recently, pre-training model has brought revolutionary changes to machine reading comprehension. The pre-training method trains the model on other related tasks at first, and then further optimizes the model on the target task to transfer the knowledge learned by the model. The most representative is Bert proposed by Google in 2018 [9]. Bert uses unsupervised learning to pre-train on large-scale corpus, and creatively uses mask design and the next text judgement to enhance the language ability of the model. In the SQuAD 2.0 competition,

经审理查明,2009年六七月份,被告人林1在未取得《医师资格证书》的情况下,受聘于被告人邵3经营的台州市椒江区枫山社区卫生服务站从事医疗活动,林1于2010年6月30日被台州市椒江区卫生局当场查获,该局于2010年9月27日作出行政处罚10000元。而后,邵3继续聘请林1在该服务站工作,期间,林1购得伪造的《医师资格证书》,但未在本市路桥区卫生局注册。2011年9月27日被告人林1被椒江区卫生局再次查获。2012年2月21日、2012年2月4日,被告人林1、邵3分别经公安机关传唤到案。案发后,被告人林1退出违法所得34500元。

问题一：被告人林1在何处从事医疗活动？	回答一：台州市椒江区枫山社区卫生服务站
问题二：被告人林1被行政处罚多少？	回答二：10000元
问题三：被告人林1是否注册《医师资格证书》？	回答三：NO
问题四：被告人林1、邵3是否被公安机关传唤？	回答四：YES
问题五：被告人邵3是否退出违法所得？	回答五：UNKNOWN

Figure 1: A Case Instance.

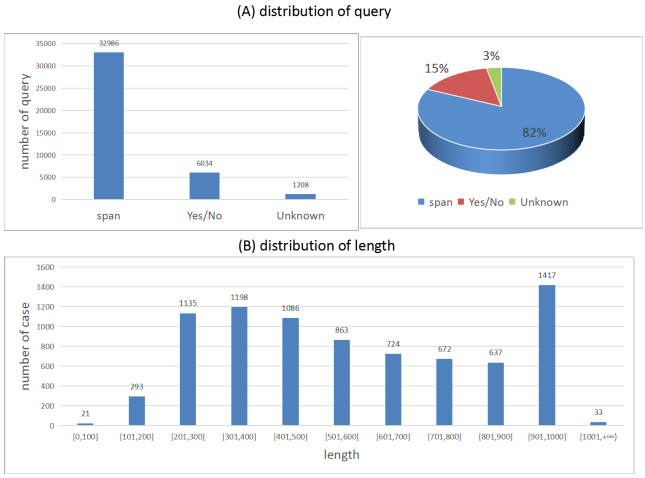

Figure 2: The Distributions of Query and Length.

the top 20 models are all based on Bert; In CoQA competition, the top 10 models are all based on bert. The best performance of the models in these two competitions has exceeded the human.

Although machine reading comprehension has made great progress, there is a lot of challenges on the reading comprehension of Chinese legal documents. Therefore, we propose a Chinese legal document reading comprehension framework [10]. Based on the Bert pre-training model, this framework optimizes the pre-training model specifically for Chinese legal documents, integrates a variety of deep learning algorithms, answers three kinds of questions that are span extraction question, yes/no question and unanswerable question. Finally, we test our framework with the real legal documents published by the Supreme People's Court.

2 DATA PREPARE AND ANALYSIS

2.1 Dataset

The dataset we used is the legal documents published by the network of China judicial documents, mainly involving civil and criminal judgments, with a total of 8079 legal documents, including 4271 civil documents and 3808 criminal documents. Each document consists of case description and several questions. There are 40228 questions in the whole data set, including 20503 questions in civil documents and 19725 questions in criminal documents [11]. The questions are divided into three categories: span extraction, yes/no and unanswerable. Span extraction are questions with clear answers in the case description, which can be answered through some spans in the case description. The question of yes or no is also a question that there is a clear answer in the case description. Yes or no is just replied. Unanswerable questions are questions that have no clear answer in the case description and cannot be answered. A data case is shown in Figure 1. The goal of our framework is to realize the correct answers to the three kinds of questions through the reading comprehension of the case context.

2.2 Legal Document Feature

We have carefully analyzed the types of problems, as shown in Figure2(A). It is found that the proportion of span extraction, yes or no and unanswerable questions in the total dataset is 82%, 15% and 3%, that is, 32986 span extraction questions, 6034 yes or no questions and 1208 unanswerable questions. There is a large gap in the number of three types of problems in the data set, which will bring serious data skew problems. Therefore, how to eliminate data skew should be considered when classifying problems. In addition, due to the existence of three different types of problems, it is necessary for the reading comprehension framework to be able to deal with different types of problems.

We analyze the length of case context, and the analysis results are shown in Figure2(B). It is found that the length of legal documents is generally long, and the average length of criminal documents is higher than that of civil documents. Proportion of case documents with more than 500 words is 53.79%, so the task belongs to the problem of long context reading comprehension. This is quite different from the English reading comprehension dataset SQuAD, because the article length of SQuAD is short, generally more than 300 words. Therefore, when designing the reading comprehension model of legal documents, we should consider the related problems brought by long context.

3 OUR FRAMEWORK

As shown in the Figure 3, our model includes the following parts: (1) the case context and queries of legal documents are sent to the data pre-processing module for format and text processing. (2) The processed context and queries improve the data quality through data enhancement. (3) The enhanced data generates word embedding through the pre-training module. (4) Query classification network contains several neural network classification models, which can classify and answer different types of questions. (5) The ensemble strategy connects multiple classification networks to enhance the ability to answer different questions. (6) In order to improve the reading comprehension ability of Chinese legal documents as much as possible, a post-processing module is designed in the framework to adjust the classification results to improve the performance of the model.

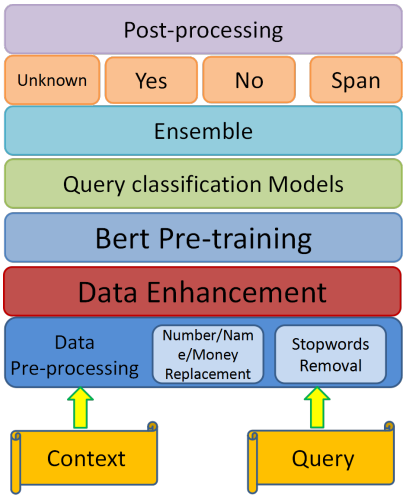

Figure 3: Architecture of Our Framework.

3.1 Data Pre-processing

In our framework, data pre-processing includes number/name/money replacement and stop words removal.

- Number/Name/Money replacement. Because some figures and money in legal documents are too accurate, which is not conducive to the positioning and matching of information by the model, we adjust the accurate figures and money to a approximate range to enhance the ability of information matching. Some persons' names add too many suffixes in privacy protection, so we also fuzzify the names of persons in a unified way to reduce mutual information interference.
- Stop words Removal. Legal documents contain many useless stop words, such as punctuation marks, modal particles, auxiliary words. Stop words are high-frequency, carry little information and needed to be removed. We use a stop words table which contains 1893 stop-use words. In stop words removal, we use the stop words table to filter legal documents, remove all the stop words, and shorten length to improve information.

3.2 Data Enhancement

The data enhancement module is mainly used to solve the significant imbalance between the three types of problems. Since the sum of yes/no questions and unanswerable questions only accounts for 18% of all questions, the model training effect of yes/no questions and unanswerable questions will be poor. For unanswerable questions, named entity recognition is used to replace the person's name and place name to increase the amount of negative question data. And exchanging the problem location to increase the amount of negative problem data is also used. For the yes/no questions, new problems are generated through the seq2seq model, and the number of positive problems is increased. For negative problems, the number of negative problems is increased through named entity recognition and manual rules. In addition, the problem of unbalanced data sets can also be solved by adjusting the answerable threshold.

3.3 Bert Pre-training

Bert (bidirectional encoder representations from transformers) [9] is a language representation model released by Google in 2018. It has successfully achieved the result of state of the art in 11 NLP tasks, and its performance exceeds human performance in some aspects. The essence of Bert is to provide a good feature representation for word learning by running the self supervised learning method on the basis of massive corpus. In a specific NLP task, we can directly use the feature representation of Bert as the word embedding of the task. Therefore, Bert provides a model of migration learning for other task. The model can be used as a feature extractor after fine tuning according to the task.

The main characteristics of Bert are as follows: (1) Transformer [12] is used as the main framework of the algorithm, and bidirectional transformer is adopted. (2) Multitasking training objectives use Mask Language Model (MLM) [13] and Next Sentence Prediction (NSP). (3) Using more powerful machines to train more large-scale data makes the results of Bert reach a new height, and Google has opened the Bert model. Users can directly use Bert as the conversion matrix of word2vec to establish a general model. For a specific task, only an additional neural network layer is needed, which is equivalent to transferring the downstream NLP work to the pre-training stage. At present, Google has open seven Bert models. We use the Chinese model (Bert base, Chinese). The model supports Chinese simplified and traditional, 12 layer, 768 hidden, 12 heads and 110m parameters.

This Chinese Bert model has a maximum length for the input sentence, which is required to be no more than 512 words. In order to solve this problem, we learn from the idea of preprocessing in Bert's fine tune scheme for SQuAD dataset. During data preprocessing, we use the sliding window method to cut the long context into multiple document spans, and for the words appearing in multiple spans, the document span with the largest context shall prevail in the subsequent calculation of scores. If answer is not in the window, set the problem to unknown. In our model, we set the window length to 512 and the number of window sliding steps to 128.

Fine tune method refers to adding a small number of task specific parameters on the basis of the trained model. For example, for classification problems, add a softmax network on the basis of the model, and then retrain on the new corpus to get better parameters. Because reading comprehension tasks of legal documents, we re-conducted fine tune on Google's original pre-training model based on all civil and criminal documents and CAIL2018 on Bert-base-Chinese released by Google, making it more suitable for reading comprehension of legal documents. Through the following experiments, it is also proved that the effect of the fine tune model is greatly improved compared with the original model published by Google.

3.4 Query Classification Models

After Bert pre-training, we employ several classical text classification models based on deep neural network to answer different kinds of queries [10]. The architecture of models in are shown in Figure 4.

HAN (Hierarchical Attention Network). HAN [14] uses word encoder, that is word level bi-directional GRU, and word level

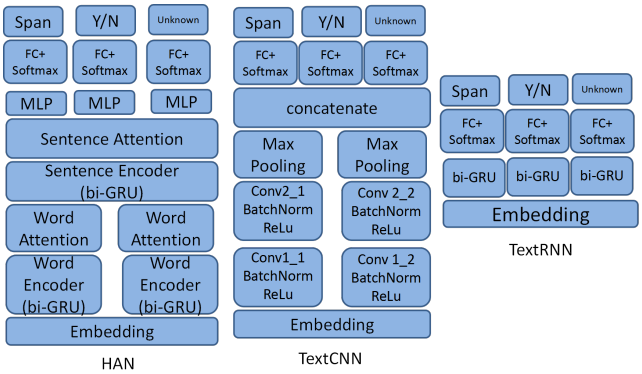

Figure 4: QueryText Classification Models in the Framework.

attention to get rich representation of words after word embedding. After word-level encoder and attention, sentence-level encoder and attention are used to get rich representation of sentences. Each MLP layer transform word representation to each queries' features. Finally fully connection and softmax function are used to compute the probability distribution.

TextCNN. In our TextCNN [15], convolutional operation, batch normalization and relu function are combined to get rich features from word embedding. Two successive combination units are employed and max pooling function gets word scalar which are concatenated to form final features. Finally, fully connection layer and softmax function will form linear layer to project these features to three queries.

TextRNN. TextRNN [16] uses bi-directional GRU layer instead of convolution layer comparing with TextCNN. Bi-directional GRU can catch uncertain length and bi-directional n-gram features, express context information better.

3.5 Ensemble Strategy

For reading and comprehension of legal documents, the ability of a single DNN model is often insufficient, and it is easy to meet bottlenecks. So how to break through the bottleneck of single model becomes a challenge. In practice, ensemble strategy is a common method to break through the bottleneck of single model.

In practice, ensemble strategy combines several different classification models with simple voting or weighted average strategies to combine their predicting results. In our framework, we provide four strategies to integrate a variety of deep learning algorithms. The details will be explained in the experiment.

3.6 Post-processing

By analyzing the answers predicted by the model, we find the following two questions:

- In criminal problems, the performance of the model is low, which is related to the complexity of criminal problems and the long answer spans.
- For some specific problems, the answer boundary predicted by the model deviates from the ground truth, which is caused by inconsistent annotation.

To solve the above problems, we add expert rules in the post-processing module to post-process the answers predicted by the model to obtain more reasonable output. The rules include using entity recognition method to verify the entity names extracted from the context, such as person name and company name, with the answer of model; matching the charges span in criminal documents; correction of the form that the model answer does not conform to the date format.

The post-processing module has strong pertinence and can be adjusted according to specific problems, so as to improve the performance of the whole framework.

4 EXPERIMENTS

4.1 Evaluation Metrics

We used macro average F to evaluate the results of reading comprehension [17]. For each query, the prediction result needs to be calculated with N standard answers to obtain N F, and the maximum value is taken as its F. In order to compare indicators more fairly, N standard answers are divided into N groups with N-1 answers. Finally, the F value of each query is the average value of F in N groups. The F value of the whole data set is the average F of all data. The calculation formula is as follows:

$$L_g = len\,(gold)\,, L_p = len\,(pred)$$

$$L_I = Intersection\,(gold, pred)$$

$$P = \frac{L_I}{L_p}\,, R = \frac{L_I}{L_g}\,, F\,(gold, pred) = \frac{2 \times P \times R}{P + R}$$

$$Average\,(F) = \frac{\sum_{i=1}^{Standard} max\,(F\,(gold, pred))}{Standard}$$

$$F_{macro} = \frac{\sum_{i=1}^{N} Average\,(F_i)}{N}$$

Intersection calculates the intersection of predicted answers and standard answers with words, standard represents the number of standard answers that is 3, and max takes the maximum value F of predicted answers and each standard answer. The final score is the average of the macro average F value of the test set.

4.2 Results

We randomly select 7000 cases including 34855 queries for training and 1079 cases including 5373 queries for testing, send context and query as input into pre-train models to reading comprehension of Chinese legal documents.

Our experiment includes three parts. The first part is to reading comprehension of legal contexts by single models. The second part is to optimize pre-training model of Bert model through fine tune so as to improve the performance of reading comprehension for Chinese legal documents. The third part is to reading comprehension of legal contexts by ensemble strategy.

In the first part of the experiment, we focus on the effect of data enhancement to reading comprehension. Therefore, we adjusted a few parameters of each model quickly. The results of reading comprehension by each model are shown in the Table 1.

As table shows, data enhancement is beneficial to improve the performance. After data enhancement, the prediction results for reading comprehension have been improved. The performance of

Table 1: The Result of Reading Comprehension of Single Models

Models	Unenhancement	Enhancement
HAN	59.881	66.382
TextCNN	55.368	62.288
TextRNN	56.249	63.95

Table 2: The Result of Reading Comprehension of Single Models after Legal Contact Pre-training

Models	F_{macro}	Improved Ratio
HAN	79.619	19.94%
TextCNN	74.227	19.17%
TextRNN	77.825	21.7%

several models is not particularly good, and there is little difference between the results. The result of HAN is better because HAN has more complex network and stronger feature expression. But its result is only about 3.8% higher than TextRNN's. It can be seen that reading comprehension is a difficult task, so it is not very good to use a single deep learning model.

Bert transfers many traditional operations done in downstream natural language processing tasks to the pre-training word embedding. After obtaining Bert word embedding, only a simple MLP or linear classifier is needed to add for word embedding. In the first experiment, we used bert-base-chinese with 12 layer, 768 hidden, 12 heads and 110m parameters officially released by Google. Therefore, in the second experiment, we employed pre- training of Bert model with legal documents similar to the Chinese legal document after enhancing the data, so as to compare the performance of pre-training of Bert with different data.

In the second experiment, we first used the documents with the same case to pre-training of Bert model in first experiment. The purpose of this pre-training is to make the pre-training model learn as much as possible the relationship between case context of documents in the test set. Then the training data of documents and query are used for pre-training again, the purpose of this pre-training is to make the pre-training model learn the relationship between the context of training data. The experiment results are shown in the Table 2. From the results of the second experiment, the F_{macro} of HAN, TextCNN, TextRNN have been separately improved 19.94%, 19.17%, 21.7% compared with enhancement models. Therefore, it shows that the performance of the model can be significantly improved by pre-training of Bert with targeted data for reading comprehension.

In the third part of the experiment, we research how to break through the bottleneck of single model by ensemble strategy.

We test four ensemble strategies. The specific strategies are as follows:

Strategy I: select the model with the largest score from multiple base models as the final prediction.

Strategy II: deal with yes/no query, unanswerable query and span query respectively, use the combination of maximum score and voting. If half of the base models predict yes/no or unanswerable, the answer is decided according to the voting; Otherwise, the final answer is decided according to the maximum score.

Table 3: The Result of Ensemble Strategy

Strategy	F_{macro}
I	79.923
II	80.37
III	82.773
IV	81.015

Strategy III: completely use voting. If a span is predicted by multiple base models, it is chosen as the answer.

Strategy IV: average the probability logits predicted by all base models as the final prediction.

The results of ensemble strategies are shown in Table 3.

It can be seen from the table that strategy III of fully voting, that is voting by the predictions of multiple base models and selecting the prediction with the most votes as the final answer, has the highest score of 82.773. In the Chinese Judicial Reading Comprehension Challenge 2019, the F_{macro} of champion is 83.386 and gap of 0.74% is existed.

5 CONCLUSION

We research the reading comprehension of Chinese legal documents. Through our research, we find that there are some problems such as difficult analysis, unbalanced data. To solve these problems, we use pre-training model based Bert to build a reading comprehension framework of Chinese legal documents. Based on the Bert pre-training model, the framework uses a variety of deep learning models by reading legal documents to answer three kinds of questions: span extraction, yes/no and unanswerable queries. In our framework, we comprehensively use data enhancement, Bert pre-training based on legal documents, ensemble strategy to solve three kinds of questions accurately.

REFERENCES

[1] Hermann K M, Tomávs Kovcisk, Grefenstette E, *et al.* (2015). Teaching Machines to Read and Comprehend[C]. Advances in Neural Information Processing Systems.
[2] Chen D, Bolton J, Manning C D. (2016). A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task[J].
[3] Rajpurkar P, Zhang J, Lopyrev K, *et al.* (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text[J].
[4] Xiong C, Zhong V, Socher R (2016). Dynamic Coattention Networks For Question Answering[J].
[5] Wang W, Nan Y, Wei F, *et al.* (2017). Gated Self-Matching Networks for Reading Comprehension and Question Answering[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
[6] Wang W, Yan M, Wu C (2018). Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering[J].
[7] Field B, Znati T F, D Mosse (2000). V-NET: A versatile network architecture for flexible delay guarantees in real-time networks[J]. IEEE Transactions on Computers, 49(8): 841-858.
[8] Wang, Y. *et al.* (2018). Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification.arXiv:1805.02220.
[9] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[10] Yuan L, Wang J, Fan S, *et al.* (2019). Automatic Legal Judgment Prediction via Large Amounts of Criminal Cases[C].2019 IEEE 5th International Conference on Computer and Communications (ICCC). IEEE.
[11] http://cail.cipsc.org.cn/

[12] Vaswani, Ashish, *et al.* (2017). Attention is all you need. Advances in Neural Information Processing Systems.
[13] Taylor W L. (1953). "Cloze Procedure": A New Tool For Measuring Readability[J]. The journalism quarterly, 30(4):415-433.
[14] Yang Z, Yang D, Dyer C, *et al.* (2016). Hierarchical Attention Networks for Document Classification[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
[15] Kim Y (2014). Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
[16] Liu P, Qiu X, Huang X (2016). Recurrent Neural Network for Text Classification with Multi-Task Learning[J].
[17] The Chinese Judicial Reading Comprehension Challenge, SMP2019-CJRC, https://conference.cipsc.org.cn/smp2019/fayanbei.html.