LBNet: A Model for Judicial Reading Comprehension

Hao Liu and Jungang Xu(✉)

Chinese Academy of Sciences University, Beijing, China
liuhao172@mails.ucas.edu.cn, xujg@ucas.ac.cn

Abstract. In this paper, a new model for judicial reading comprehension called LBNet that combines an end-to-end network with a BERT structure is proposed, which aims to answer questions from a given passage in judicial files. Firstly, BERT is used to extract the representation of the passage and the question, and. self-matching attention mechanism is introduced to refine the representation by matching the passage against itself, which can effectively encode information from the whole passage. In the question and answer model, the pointer networks is used to locate the positions of answers from the passages. Experimental results on the CAIL2019 datasets (Chinese Judicial Reading Comprehension), show that our model can achieve good results.

Keywords: Question answering task · BERT · Judicial reading comprehension

1 Introduction

Machine reading comprehension (MRC) is a frontier field in natural language processing (NLP), which requires that machine can read, understand, and answer questions about a text. Benefiting from the rapid development of deep learning techniques (Hermann et al. 2015; Rajpurkar et al. 2016), the end-to-end neural methods have achieved promising results on MRC task (Seo et al. 2016; Huang et al. 2017; Chen et al. 2016; Clark and Gardner 2017; Hu et al. 2017; Devlin et al. 2018; Rajpurkar et al. 2018). LSTM, CNN, and attention mechanism is the common structures used in MRC. With the introduction of a series of larger and more systematic text representation models, such Bidirectional Encoder Representation from Transformers (BERT), the status of sequence representation models has been challenged. compared with the sequence representation model, a better understanding of semantics and adequate training of the article are the advantages of the pre-training model. After pre-training, simple fine-tuning can handle the problem that the sequence representation model takes a lot of time to solve. In this paper, we combine the BERT and end-to-end network models and apply them to the question and answer task.

SQuAD (Rajpurkar et al. 2018), Dureader (He et al. 2017), CoQA (Reddy et al. 2019), are the large-scale and different datasets for reading comprehension, which requires to answer questions given a passage. And in addition to the general types, the prospects for specific industry applications are now very well. In this paper, we focus on the CAIL

Q. Chen and J. Li (Eds.): APWeb-WAIM 2020 Workshops, CCIS 1373, pp. 144–155, 2021.
https://doi.org/10.1007/978-981-16-0479-9_12

(Xiao et al. 2018) datasets (Chinese Judicial Reading Comprehension dataset), The law is closely related to people's daily lives. Almost every country in the world has laws. Everyone must abide by the laws to enjoy their rights and perform their duties. Every day, tens of thousands of traffic accidents, private loans, and divorce disputes occur. At the same time, in the process of handling these cases, many judgments will be made. The verdict is usually a summary of the entire case, involving the description of the event, the opinion of the court, the result of the verdict, etc. However, there are relatively few legal staff and factors such as uneven judges can often lead to wrong decisions. Moreover, even in similar cases, the judgment results can sometimes be very different. In addition, a large number of documents makes extracting information from them extremely challenging. Therefore, introducing artificial intelligence into the legal field will help judges make better decisions and work more effectively. CAIL requires to answer questions given a civil and criminal judgment documents. The referee documents contain a wealth of case information, such as time, place, relationship, etc., through the intelligent reading and understanding of the judgment documents, the results can help judges, lawyers and the general public to obtain the required information more quickly and conveniently. This dataset is the first reading comprehension dataset based on Chinese judgment documents, which belongs to the Span-Extraction Machine Reading Comprehension. In order to increase the diversity of questions, refer to the SQuAD and CoQA. This dataset adds unanswerable and YES/NO problem. In view of the fact that the civil and criminal judgment documents differ greatly in the factual description, the corresponding types of questions are not the same. In order to take into account the two types of judgment documents at the same time, CAIL dataset will set up civil and criminal test set. An example of CAIL dataset is shown in Fig. 1.

经审查表明，原、被告于2010年11月5日登记结婚，婚生子王凯翔（现改名为那8）于2012年2月10日出生。2013年1月14日，经本院主持调节，双方当事人就抚养权自愿达成如下协议：婚生子王凯翔由葛某抚养，王某从2013年1月起每月承担抚养费10000元至王翔凯独立生活之日止。2012年9月22日，原告王某与他人孕育一男孩，离婚后原、被告均已重组家庭，现原、被告双方都有稳定的工资收入，原告王某之妻没有固定收入来源，孕育两个男孩，居住于原告王某父母房屋，被告葛某之夫有收入来源，带有一女，一家住XXXX，现就读于准格尔旗民族幼儿园。以上事实由原、被告陈述及原告出示的工资收入及存款证明、葛某、那8常驻人口登记卡、出生医学证明在案予以证实。

According to the review, the plaintiff and the defendant were registered to marry on November 5, 2010. The married son Wang Kaixiang (now renamed Na ba) was born on February 10, 2012. On January 14, 2013, after the host presided over the adjustment, the two parties voluntarily reached the following agreement on custody: Wang Kaixiang, a legitimate child, was raised by Ge, and Wang took 10,000 yuan a month from January 2013 until Wang Xiangkai Living independently. On September 22, 2012, the plaintiff Wang and other women gave birth to a boy. After the divorce, both the plaintiff and the defendant had reorganized the family. Both the plaintiff and the defendant have stable wage income. The wife of the plaintiff, Wang, has no fixed source of income and raises two boys. They live in the plaintiff's parents' house. The defendant Ge's husband has a source of income and a daughter. Now living in XXXX, her daughter is currently enrolled in the Zhungeer Banner National Kindergarten. The above facts are stated by the plaintiff and the defendant. The wage income and deposit certificate presented by the plaintiff, the resident resident registration card of Ge and Na ba and the birth medical certificate were confirmed on the case.

Q1:The two sides agree on how much money Wang pays each month.	A1:February 10, 2012
Q2:Where is the stepdaughter of the defendant Ge Mou currently studying?	A2:10,000 yuan
Q3:When is the date of birth of Wang Kaixuan?	A3:Zhungeer Banner National Kindergarten
Q4:Whether the plaintiff and the defendant respectively formed a new family	A4:Yes
Q5:Does Wang Kaixuan have the will to live with the plaintiff Wang?	A5:Unk

Fig. 1. An example item from dataset CAIL.

To understand the properties of CAIL, we analyze the questions and answers in the development set. Specifically, we explore the numbers of two types of judgment documents, and the proportion of different answer types, and distribution of documents length (Figs. 2 and 3).

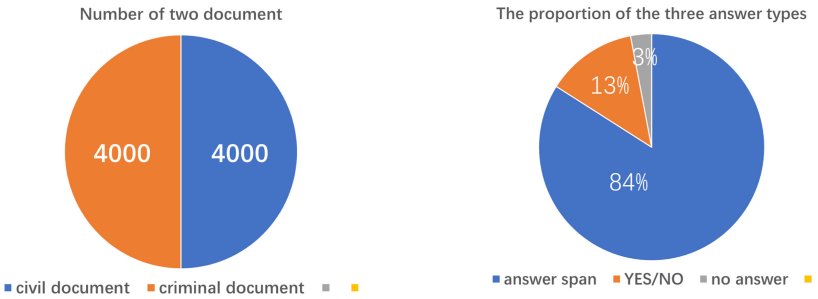

Fig. 2. Analysis of the data set

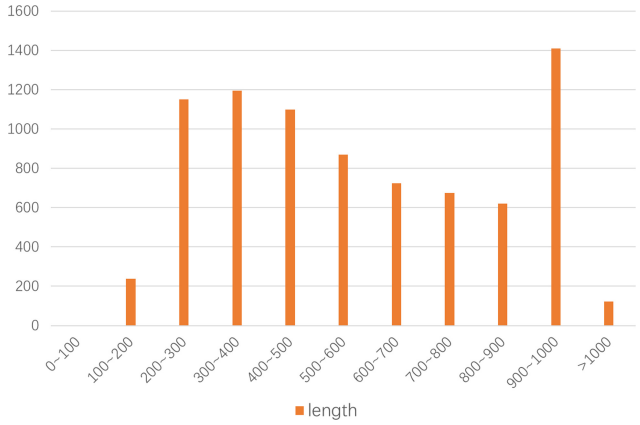

Fig. 3 Distribution of document length

The composition of the CAIL training set is mainly segment extraction, which also contains 13% of YES/NO questions and 3% of questions that cannot be answered. A reasonable solution is needed to deal with different types of questions. The length of CAIL documents is generally longer, more than 50% of the documents are longer than 500, and the long-text related issues should be considered in the model design.

To do well on MRC with unanswerable questions, the model needs to comprehend the question, reason among the passage, judge the unanswerability and then identify the answer span. When the question is answerable, the main challenge of this task lies in how to reliably determine whether a question is not answerable from the passage.

There are two kinds of approaches to model the answerability of a question. One approach is to directly extend previous MRC models by introducing a no-answer score to

the score vector of the answer span (Levy et al. 2017; Clark and Gardner 2017). But this kind of approaches is relatively simple and cannot effectively model the answerability of a question. Another approach introduces an answer verifier to determine whether the question is unanswerable (Hu et al. 2018; Tan et al. 2018). However, this kind of approaches usually has a pipeline structure. The answer pointer and answer verifier have their respective models, which are trained separately. Intuitively, it is unnecessary since the underlying comprehension and reasoning of language for these components is the same.

In this paper, we divide the questions into three categories, the answerable question, and the unanswerable question, the YES/NO question. If the question is judged to be YES/NO, it is turned into a classification question. Otherwise, first judge whether it can answer, if possible, give the start point and end point.

We propose a model called LBNet (Long-term recurrent attention network from Bert) to incorporate these three sub-tasks into a unified model: (1) an answer pointer to predict a candidate answer span for a question; (2) a no-answer pointer to avoid selecting any text span when a question has no answer; and (3) an answer verifier to determine the probability of the "YES/NO" of a question with candidate answer information Our experimental results on the CAIL dataset show that LBNet effectively predicts the unanswerability of questions and achieves an F1 score of 83.5.

2 LBNet Model

For reading comprehension style question answering, a passage P and question Q are given, our task is to predict an answer A to question Q based on information found in P. The CAIL dataset further constrains answer A either to be a continuous sub-span of passage P or is YES/NO. Answer A often includes non-entities and can be much longer phrases. This setup challenges us to understand and reason about both the question and passage in order to infer the answer.

The BERT model is based on the powerful model of Transformer, which itself has broken the record of many natural language processing directions created by the deep neural network model. In general, it has been able to deal with many problems and achieve good results, However, the traditional long short-term memory network also has its advantages because that it can handle the contextual relationship well and retain the key information. So, people wish to achieve better results by combining these two models. Therefore, we made some changes based on the original BERT model and explored a new model, called LBNet (Long-term Recurrent Integrate BERT Network), which can handle machine reading task better.

LBNet is a contextual attention-based deep neural network for the task of conversational question answering, in which, the bottom layer is the input vector, and it is constructed in the same way as BERT, which is a combination of Position Embeddings, Token Embeddings and Segment Embeddings. LBNet has similar stems with existing machine reading comprehension models, but it also has several unique characteristics to tackle contextual understanding during conversation. Firstly, LBNet applies self-attention on passage and question to obtain a more effective understanding of the passage and dialogue history. Secondly, LBNet leverages the latest breakthrough in

BERT contextual embedding (Devlin et al. 2018). Different from the canonical way of appending a thin layer after BERT structure according to (Devlin et al. 2018), we innovatively employed the BiLSTM layer outputs, with locked BERT parameters. Empirical results show that each of these components has substantial gains in prediction accuracy. An illustration of LBNet model is shown in Fig. 4.

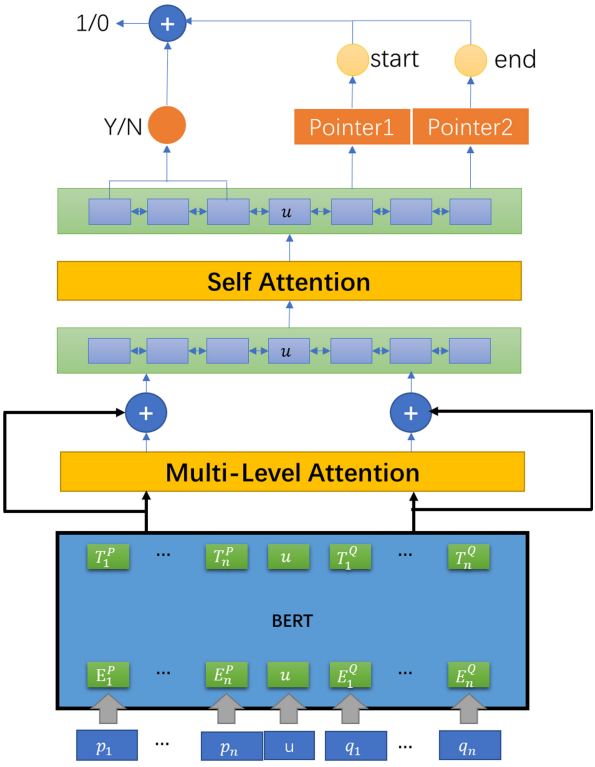

Fig. 4. LBNet Model for CAIL datasets

Formally, we can represent the MRC problem as: given a set of tuples (Q, P, A), where $Q = (q_1, q_2, \ldots, q_m)$ is the question with m words, $P = (p_1, p_2, \ldots, p_n)$ is the context passage with n words, and $A = p_{(r_s):(r_e)}$ is the answer with r_s and r_e indicating the start and end points, the task is to estimate the conditional probability P (A|Q, P), LBNet consists of four major blocks: Bert & BiLSTM Encoding, Multi-Level Attention, Final Fusion, and Prediction.

We first combine the embedded representation of the question and passage with a universal node u and pass them through a Bert and BiLSTM to encode the whole text. We then use the encoded representation to deal the information interaction. Then we use the encoded and interacted Representation to fuse the full representation and feed them into the final prediction layers to do conduct the prediction. We will describe our model in details in the following.

2.1 BERT and BiLSTM Encoding

Embedding

We first segment Chinese sentences into words. Then embed both the question and the passage with the following features. Glove embedding (Pennington et al. 2014) and Elmo embedding (Peters et al. 2018) are used as basic embeddings. Besides, we use POS embedding and NER embedding (Luo et al. 2019), we use 12 dimensions to embed POS tags, 8 for NER tags, and a feature embedding that includes the exact match, lowercase match, lemma match, and a TF-IDF feature. Now we split the question Q into $Q = \left\{w_t^Q\right\}_{t=1}^m$, and the passage P into $P = \left\{w_t^P\right\}_{t=1}^n$.

Consider the question $Q = \left\{w_t^Q\right\}_{t=1}^m$ and the passage $P = \left\{w_t^P\right\}_{t=1}^n$. We first convert the words to their respective word-level embeddings ($\left\{e_t^Q\right\}_{t=1}^m$ and $\left\{e_t^P\right\}_{t=1}^n$) and character-level embeddings ($\left\{c_t^Q\right\}_{t=1}^m$ and $\left\{c_t^P\right\}_{t=1}^n$). The character-level embeddings are generated by taking the final hidden states of a bi-directional recurrent neural network (RNN) applied to embeddings of characters in the token. And E_Q denotes Q's segment embeddings. E_P denotes P's segment embeddings. E_i^{m+n+1} denotes position embeddings. The input embeddings is the sum of the token embeddings (word-level and character-level), the segment embeddings and the position embeddings. Now we get the question representation $Q = q_{i=1}^m$ and the passage representation $P = p_{i=1}^n$, where each word is represented as a d-dim embedding by combining the features/embedding described above.

The universal node u is first represented by a d-dim randomly-initialized vector. The universal node u can connect passage and questions. We concatenated question representation Q, universal node representation u, passage representation P together as:

$$V = [Q, u, P] = \left[q_1, q_2 \ldots q_m, u, p_1, p_2, \ldots, p_n\right] \tag{1}$$

$V \in \mathbb{R}^{d\times(m+n+1)}$ is a joint representation of question universal node, and passage.

Word-Level Fusion

Then we first use Bert model (Devlin et al. 2018) and bidirectional LSTM (BiLSTM) to fuse the joint representation of question, universal node, and passage.

$$H^l = \text{Bert}(V) \tag{2}$$

And we pass it through the third BiLSTM and obtain a full representation H^f

$$H^f = \text{BiLSTM}\left(H^l\right) \tag{3}$$

We concatenate H^l and H^f together, Thus, $H = \left[H^l; H^f\right]$ represents the deep fusion information of the question and passage on word-level. When a BiLSTM is applied to encode representations, it can learns the semantic information bi-directionally.

2.2 Multi-level Attention

To fully fuse the semantic representation of the question and passage, we use the attention mechanism (Bahdanau et al. 2014) to capture their interactions on different levels.

We first divide H into two representations: attached passage H_q and attached question H_p, and let the universal node representation h_{m+1} attached to both the passage and question, i.e.

$$H_q = \left[h_1, h_2, \ldots, h_{m+1}\right] \tag{4}$$

$$H_p = \left[h_{m+1}, h_{m+2}, \ldots, h_{m+n+1}\right] \tag{5}$$

Since both $H_q = \left[H_q^l, H_q^f\right]$ and $H_p = \left[H_p^l, H_p^f\right]$ are concatenated by three-level representations, we followed previous work FusionNet (Huang et al. 2017) to construct their iterations on three levels. Take the first level as an example. We first compute the affine matrix of H_l^q and H_p^l by

$$S = \left(\mathrm{ReLU}\left(W_1 H_q^l\right)\right)^T ReLU\left(W_2 H_p^f\right) \tag{6}$$

where $S \in \mathbb{R}^{(m+1)\times(n+1)}$; W_1 and W_2 are learnable parameters. Next, a bi-directional attention is used to compute the interacted representation $\widetilde{H_q^l}$ and $\widetilde{H_p^l}$.

$$\widetilde{H_q^l} = H_q^l \times \mathrm{softmax}\left(S^T\right) \tag{7}$$

$$\widetilde{H_p^f} = H_q^l \times \mathrm{softmax}(S) \tag{8}$$

where softmax(·) is column-wise normalized function. We use the same attention layer to model the interactions for all the three levels, and get the final fused representation $\widetilde{H_q^l}$, $\widetilde{H_p^f}$ for the question and passage respectively.

2.3 Final Fusion

After the three-level attentive interaction, we generate the final fused information for the question and passage. Following the work of Sun (2018), we concatenate all the history information: we first concatenate the encoded representation H and the representation after attention $\tilde{H}$ (again, we use H^l, H^f, and $\widetilde{H^l}$, $\widetilde{H^f}$ to represent two different levels of representation for the two previous steps respectively).

First, we pass the concatenated representation H through a BiLSTM to get H^A.

$$H^A = \mathrm{BiLSTM}\left(\left[H^l; H^f; \widetilde{H^l}; \widetilde{H^f}\right]\right) \tag{9}$$

where the representation H^A is a fusion of information from different levels.

Then we concatenate the original embedded representation V and H^A for better representation of the fused information of passage, universal node, and question

$$A = \left[V; H^A\right] \tag{10}$$

Finally, we use a self-attention layer to get the attention information within the fused information.

$$\tilde{\mathrm{A}} = A \times \mathrm{softmax}\left(\mathrm{A}^{\mathrm{T}}\mathrm{A}\right) \tag{11}$$

Next we concatenate H^{A} and $\tilde{\mathrm{A}}$ and pass them through another BiLSTM layer.

$$O = \mathrm{BiLSTM}\left[\mathrm{H}^{\mathrm{A}}; \tilde{A}\right] \tag{12}$$

We divide O into two parts: O^{P}, O^{Q}, which denote the fused information of the question and passage respectively

$$\mathrm{O}^{\mathrm{P}} = [\mathrm{o}_1; \mathrm{o}_2; \ldots; \mathrm{o}_{\mathrm{m}}] \tag{13}$$

$$\mathrm{O}^{\mathrm{Q}} = \left[\mathrm{o}_{\mathrm{m}+1}; \mathrm{o}_{\mathrm{m}+2}; \ldots; \mathrm{o}_{\mathrm{m}+\mathrm{n}+1}\right] \tag{14}$$

2.4 Prediction

We follow the work of Wang and Jiang (2015) and use pointer networks (Vinyals et al. 2015) to predict the start and end position of the answer.

First, we use a function shown below to summarize the question information O^{Q} into a fixed-dim representation c_{q}.

$$\mathrm{c}_{\mathrm{q}} = \frac{\exp\left(\mathrm{W}^{\mathrm{T}}\mathrm{o}_{\mathrm{i}}^{\mathrm{Q}}\right)}{\sum_{\mathrm{j}} \exp\left(\mathrm{W}^{\mathrm{T}}\mathrm{o}_{\mathrm{j}}^{\mathrm{Q}}\right)} \mathrm{o}_{\mathrm{i}}^{\mathrm{Q}} \tag{15}$$

We use two trainable matrices W_{s} and W_{e} to estimate the probability of the answer start and end boundaries of the i_{th} word in the passage, α_{i} and β_{i}.

$$\alpha_i \propto \exp\left(\mathrm{c}_{\mathrm{q}}\mathrm{W}_{\mathrm{s}}\mathrm{o}_{\mathrm{i}}^{\mathrm{p}}\right) \tag{16}$$

$$\beta_i \propto \exp\left(\mathrm{c}_{\mathrm{q}}\mathrm{W}_{\mathrm{e}}\mathrm{o}_{\mathrm{i}}^{\mathrm{P}}\right) \tag{17}$$

And we use the weight matrix obtained from the answer pointer to get two representations of the passage.

$$\mathrm{c}_{\mathrm{s}} = \sum_{\mathrm{i}} \alpha_{\mathrm{i}} \cdot \mathrm{o}_{\mathrm{i}}^{\mathrm{p}} \tag{18}$$

$$\mathrm{c}_{\mathrm{e}} = \sum_{\mathrm{i}} \beta_{\mathrm{i}} \cdot \mathrm{o}_{\mathrm{i}}^{\mathrm{p}} \tag{19}$$

To train the network, we minimize the sum of the negative log probabilities of the ground truth start and end position by the predicted distributions.

3 Experiment

3.1 Dataset

The dataset used in the technical evaluation of this task is provided by HKUST Xunfei. The dataset mainly comes from the referee documents of China Referee Documents Network, which includes criminal and civil first instance referee documents.

The training set contains about 40,000 questions, and the development set and test set each have about 5000 questions respectively. For the development set and the test set, each question contains 3 manually labeled reference answers.

In view of the large differences in the factual description of the civil and criminal adjudication documents, and the corresponding types of questions are not the same, in order to take into account both types of adjudication documents at the same time, thereby covering most of the adjudication documents, they are divided into civil and criminal test sets.

3.2 Metrics

This task is evaluated using a macro-average F1 that is consistent with the CoQA competition. For each question, need to be calculated with N standard answers to get N F1 scores, and the maximum value is taken as its F1 value. However, in assessing Human Performance, each standard answer requires an F1 value to be calculated with N-1 other criteria. In order to compare indicators more fairly, N standard responses need to be divided into N groups according to the N-1 group. Finally, the F1 value of each problem is the average of the N groups F1. The F1 value of the entire data set is the average of all data F1. The F1 value of the entire data set is the average of all data F1.

$$Lg = len(gold) \tag{20}$$

$$Lp = len(pred) \tag{21}$$

$$Lc = InterSec(gold, pred) \tag{22}$$

$$precision = \frac{Lc}{Lp} \tag{23}$$

$$recall = \frac{Lc}{Lg} \tag{24}$$

$$f1(gold, pred) = \frac{2 \times precision \times recall}{precision + recall} \tag{25}$$

$$Avef1 = \frac{\sum_{i=1}^{count_{ref}} \max(f1(gold_i, pred))}{Count_{ref}} \tag{26}$$

$$F1_{macro} = \frac{\sum_{i=1}^{N} Avef1_i}{N} \tag{27}$$

InterSec calculates the intersection of the predicted answer and the standard answer (in words), Countref represents the number of standard answers (three), max part takes the predicted answer and each standard answer, the maximum value of the F1 value. The final score is the average of the average F1 values for the criminal and civil test sets.

3.3 Implementation Details

We use Spacy to process each question and passage to obtain tokens, POS tags and NER tags of each text. We use 10 dimensions to embed POS tags, 10 for NER tags (Luo et al. 2019). We use 100-dim Glove pretrained word embeddings and 1024-dim Elmo embeddings. All the LSTM blocks are bi-directional with one single layer. We set the hidden layer dimension as 125, attention layer dimension as 250. We added a dropout layer over all the modeling layers, including the embedding layer, at a dropout rate of 0.3. We use Adam optimizer with a learning rate of 0.002.

3.4 Experimental Results and Analysis

Baseline Moels and Metrics

We compare LBNet with the following baseline models: LibSVM (Chang et al. 2011), BiDAF (Seo et al. 2016), (Devlin et al. 2018), ERNIE (Zhang et al. 2019). The dataset is randomly partitioned into a training set (80%), a development set (20%). We use F1 as the evaluation metric, which is the harmonic mean of precision and recall at word level between the predicted answer and ground truth.

4 Results

Table 1 shows the experimental results of LBNet and baseline models on CAIL datasets. As shown in Table 1, LBNet achieves better results than all baseline models. In detail, LBNet model improves F1 by 19.8, 16.4, 7.8, 6.7 on civil dataset and 17.3, 14.8, 7, 4.2 on criminal dataset compared with LibSVM, BiDAF, ERNIE and BERT, respectively. To be noted that we use the pretrain model of BERT and ENGIE. BERT uses MLM (Masked Language Model) to obtain context-relevant bidirectional feature representations. ENRIE introduces knowledge, combining entity vectors with textual representation. Different from the previous models, we use a unified representation to encode the question and passage simultaneously, and introduce a universal node which plays an important role to predict the unanswerability of a question, and we use the BiLSTM for encoding the embedded representation, which is very effective to fuse information of the question and passage.

Table 1. Experimental results (F1) on the CAIL dataset

Model	Civil data set	Criminal data set
LibSVM	63.5	63.8
BiDAF	66.8	66.5
ERNIE	75.2	74.3
BERT	78.2	77.3
LBNet	82.9	80.9

5 Conclusions

In this paper, we propose a novel contextual attention-based model, LBNet, to tackle Judicial Reading Comprehension tasks. For the joint learning of different types of questions to design an "answer fragment extraction" and "YES/NO classification and unanswerable question" three tasks of the end-to-end model, the different types of problems unified learning. For the long text problem, draw on the idea of pre-processing in the fine-tune solution for the SQuAD dataset, which is to use the sliding window method to cut the long text into multiple doc_span when data is preprocessed, for words that appear in multiple spans, the doc_span of the word with "maximum context" prevails when the score is subsequently calculated. Following an in-depth analysis of the data set, we found that some of the problems have some laws or the answers to the model prediction can be further corrected, so the post-processing module was added to the overall model structure to further improve performance. By leveraging inter-attention and self-attention and using BiLSTM on passage and conversation history, the model is able to comprehend dialogue flow and fuse It with the digestion of passage content. Furthermore, we incorporate the latest breakthrough in NLP, BERT, and leverage it in an innovative way. LBNet achieves good results over previous approaches. On the dataset CAIL, LBNet achieves F1 score 83.5 and 81.3 accuracy. In the future, we will further optimize the network structure and parameters to get more accurate results.

References

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)

Hermann, K.M., et al.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems (2015)

Reddy, S., Chen, D., Manning, C.D.: CoQA: A conversational question answering challenge. Trans. Assoc. Comput. Linguist. **7**, 249–266 (2019)

Huang, H.Y., Zhu, C., Shen, Y., Chen, W.: Fusionnet: Fusing via fully-aware attention with application to machine comprehension. arXiv preprint arXiv:1711.07341 (2017)

Clark, C., Gardner, M.: Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723 (2017)

He, W., et al.: Dureader: a chinese machine reading comprehension dataset from real-world applications. arXiv preprint arXiv:1711.05073 (2017)

Hu, M., Peng, Y., Huang, Z., Qiu, X., Wei, F., Zhou, M.: Reinforced mnemonic reader for machine reading comprehension. arXiv preprint arXiv:1705.02798 (2017)

Xiao, C., Zhong, H., Guo, Z., et al.: CAIL2018: a large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)

Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822 (2018)

Luo, R., Xu, J., Zhang, Y., et al.: PKUSEG: a toolkit for multi-domain Chinese word segmentation. arXiv preprint arXiv:1906.11455 (2019)

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

Wang, S., Jiang, J.: Learning natural language inference with LSTM. arXiv preprint arXiv:1512.08849 (2015)

Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems, pp. 2692–2700 (2015)

Sun, F., Li, L., Qiu, X., Liu, Y.: U-Net: machine reading comprehension with unanswerable questions. arXiv preprint arXiv:1810.06638 (2018)

Dhingra, B., Yang, Z., Cohen, W.W., Salakhutdinov, R.: Linguistic knowledge as memory for recurrent neural networks. arXiv preprint arXiv:1703.02620 (2017)

Dhingra, B., Liu, H., Yang, Z., Cohen, W.W., Salakhutdinov, R.: Gated-attention readers for text comprehension. arXiv preprint arXiv:1606.01549 (2016)

Chen, D., Bolton, J., Manning, C. D.: A thorough examination of the CNN/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858 (2016)

Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)

Sun, F., Li, L., Qiu, X.: U-Net: machine reading comprehension with unanswerable questions. arXiv preprint arXiv:1810.06638 (2018)

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)