

A novel knowledge enhancement method for large-scale natural language training model

Qi Han^{1,3}, Gilja So^{2*}

¹Department of Computer and Information Engineering, Graduate School Youngsan University, South Korea

²Department of Cyber Security Youngsan University, South Korea

³Artificial Intelligence College, Shenyang Normal University, Shenyang, 110034, China

Abstract

Knowledge enhancement-based large-scale natural language training model is an advanced language model that combines deep learning and knowledge enhancement. By learning from massive unlabeled data and combining with external knowledge such as knowledge graph, it breaks through the limitations of traditional models in interpretability and reasoning ability. Introducing knowledge into data-driven artificial intelligence model is an important way to realize human-machine hybrid intelligence. However, since most pre-trained models are trained on large-scale unstructured corpus data, the defects in certainty and explainability can be remedied to some extent by introducing external knowledge. To solve the above problems, we present a knowledge-enhanced large-scale natural language training model that integrates deep learning with external knowledge sources (e.g., knowledge graphs) to improve interpretability and reasoning ability. This approach addresses the limitations of traditional models trained on unstructured data by incorporating external knowledge to enhance certainty and explainability. We propose a new knowledge enhancement method and demonstrate its effectiveness through a long text representation model. This model processes structured, knowledge-rich long texts by extracting and integrating knowledge and semantic information at the sentence and document levels. It then fuses these representations to generate an enhanced long text representation. Experiments on legal case matching tasks show that our model significantly outperforms existing methods, highlighting its innovation and practical value.

Keywords: large-scale natural language training model, knowledge enhancement, long text representation, pre-trained model

Received on 29 March 2025, accepted on 05 July 2025, published on 15 July 2025

Copyright © 2025 Qi Han *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](#), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/airo.8987

1. Introduction

Data-driven and knowledge-driven are two important approaches to realize artificial intelligence. In recent years, data-driven methods represented by deep learning have achieved great success on various tasks. In the field of natural language processing, pre-trained language models trained on large-scale corpus show good performance on various tasks [1,2]. However, the current data-driven approach still has major limitations. Ahmad et al. [3] believed that its application scenarios were limited to those with sufficient knowledge or data, stability, complete information, static, domain-specific and single tasks. The

knowledge-driven artificial intelligence has good logic and interpretability, but it relies heavily on the knowledge and rules defined by humans, lacks the ability to abstract and learn features. It is difficult to represent human experience and knowledge completely [4,5].

The large language model (LLM) has made great progress, and many efficient models have been proposed, such as ChatGPT, LLaMAP, and the Chinese model BaiChuan [6,7]. These models perform well in various fields, reflecting the breadth of their expressive power. In the vertical domain, pre-training and fine-tuning on the base model through the domain data set can make the large model acquire the dialogue ability in the vertical domain. However, in the vertical field, although LLM has been fine-tuned, due to the lack of large-scale training of professional

*Corresponding author. Email: kjso@ysu.ac.kr

knowledge in the corresponding field, the understanding of professional knowledge is poor, and its expression ability is severely restricted [8].

In the professional area, large models can generate incorrect answers and reasoning, and even "hallucination" phenomena. In addition, the model trained by fine-tuning the data set may produce "forgetting" phenomenon, which may cause the model to lose the original dialogue generation ability [9]. Certain key industries, such as medicine, have higher requirements for model accuracy and safety than common fields. On the basis of maintaining the original dialogue ability of the big model, improving the accuracy and reliability of the big model's answer in the professional field has become an important issue in the vertical application of the big model.

The challenge of injecting external knowledge into long-text representations lies in effectively integrating structured knowledge with the complex semantics of long texts while preserving their hierarchical structure and contextual nuances. Specifically, long texts often contain rich, domain-specific information and intricate relationships that require careful alignment with external knowledge sources. This process must ensure that the knowledge is accurately mapped to relevant text segments, and that the resulting representation enhances interpretability and reasoning without compromising the original text's meaning. Additionally, the computational complexity of processing long texts and aligning them with knowledge graphs or other external knowledge bases can be significant, requiring efficient algorithms and scalable solutions.

By integrating knowledge into the deep learning model, the generalization ability of the model can be improved to a certain extent, and the controllability of the model can be enhanced. At present, the relevant work of knowledge extraction and knowledge graph construction from large-scale data has gradually matured, but due to the heterogeneity of knowledge and training corpus or data, the methods of knowledge empowering deep learning models and guiding application practice are still insufficient. Therefore, by summarizing and analyzing the knowledge enhancement methods of natural language pre-training models, this paper presents the approaches and development trends of knowledge enhancement of deep learning models, and provides referential ideas for the realization of general human-machine hybrid intelligence. Our main contributions are as follows.

(1) To demonstrate the effect of the proposed knowledge enhancement method in the large-scale natural language training model, we take the long text representation model as an example. In reality, there are a lot of long text data with rigorous structure and professional knowledge background. The long text is cut by sentence, and the knowledge information and semantic information are extracted sentence by sentence.

(2) Then the knowledge information and semantic information are processed differently according to the document hierarchy, and the semantic representation and knowledge representation at document level are obtained

respectively. A long text representation model is constructed which integrates knowledge information and structure information.

(3) Finally, the proposed model fuses two document-level representations to get the final long text representation. The matching experiments of similar cases under the background of legal knowledge show that the long text representation model based on knowledge enhancement is effective.

2. Related works

Pre-trained language models are divided into two stages: pre-trained word embeddings (PWE) and pre-trained contextual encoders (PCE). PWE (also known as word vector) technology is a method to embed word semantics into low-dimensional, dense and fixed-length numerical vectors using large-scale text corpus [10]. Its construction process mainly follows the word distribution hypothesis and word co-occurrence statistics [11]. Before word embedding technology is widely used, the feature vector of text is usually initialized randomly, that is, assuming that there is "enough" training data, the feature vector of a word can be considered as a parameter of the model, and it can be adjusted to a suitable vector representation through the training of the neural network [12]. For tasks with insufficient training data, word embedding technology can be regarded as an effective method to introduce external knowledge, the source of which is large-scale corpus.

Arisoy et al. [13] proposed the NNLM model, but due to the complexity of the model and the level of computing power at that time, NNLM was not widely used. Choudhary et al. [14] proposed skip-gram and CBOW models (also known as Word2Vec model) by optimizing the NNLM model. Word2Vec could not only be quickly trained on large-scale corpus, but also well represent the semantics of words [15]. Since then, models such as GloVe [16] based on global corpus and word co-occurrence matrix, and FastText [17] which added word n-gram information also appeared one after another. Although pre-trained word embedding has achieved great success, it still has the following shortcomings:

(1) Since the same word in static word embedding only corresponds to one embedded representation, the representation of the word cannot be adjusted according to the context, and it is difficult to solve the polysemy problem of the word.

(2) Shallow neural networks are not sufficient to capture complex information in large-scale corpus, and because the pre-trained network structure is not reused, the dependency between words cannot be transmitted to downstream tasks.

(3) Some rare words and unknown words are not sufficiently trained, and there are errors in the representation of these words.

When the model can not judge the meaning of the text from the context, it needs to further introduce the related attributes, states and other background knowledge of the entities involved, or domain limit the text. Ma et al. [18]

used cloze to test the ability of pre-trained models in knowledge storage. The test took triad information from knowledge graphs (including Google-RE, T-Rex, ConceptNet, etc.), and transformed it into a cloze form for the pre-trained model to make prediction. The experiment showed that when BERT gave 10 candidate answers, the hit rate was close to 60%, and the hit rate of the first 100 candidate answers was close to 80%, which was not ideal when only one answer was given, but it could already show that the pre-trained model could store some relatively general knowledge well. At the same time, BERT performs well in one-to-one relationship prediction, but poorly in many-to-many relationship prediction.

Kamalloo et al. [19] performed open-domain question-answering tasks (Natural Questions, WebQuestions, and TriviaQA) without retrieving any external context or knowledge by fine-tuning the pre-trained model T5. Experiments showed that with the increase of training scale, the effect of T5 model was also improved, even with the most advanced retrieval model.

The pre-trained word embedding model and the pre-trained context encoder have achieved great success in various natural language processing tasks by learning on large-scale corpus. However, limited by the size of the corpus, the long tail phenomenon and the learning ability of the model, the current language models still have the problem of lack of knowledge [20,21]. The knowledge enhancement studied in this paper refers to the relevant methods to improve the shortcomings of the model and improve the performance of the model by introducing artificial knowledge information.

Most current language models are divided into two stages, a pre-training stage that is task-agnostic and a task-specific stage that is task-specific. Therefore, the basic modes of knowledge enhancement for language models can also be divided into two categories as shown in Figure 1. First, knowledge is introduced in the pre-training stage; The second is the introduction of knowledge at the task-related stage. These two modes of knowledge enhancement are similar to human learning habits.

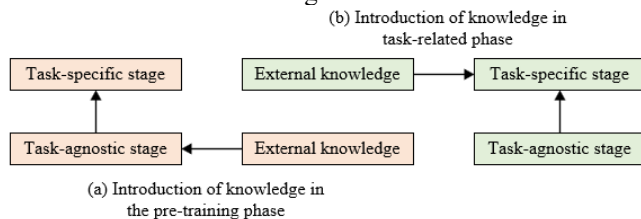


Figure 1. Two basic modes of knowledge enhancement in language models

The knowledge introduced in the task-independent pre-training stage usually covers a wide range, so that the model can better obtain the global knowledge in one or several knowledge bases. The knowledge introduced in the

task-related stage is usually closely related to specific texts and tasks and involves knowledge retrieval, so the introduced knowledge is more accurate, but it is also easy to be limited by local knowledge.

Current pre-trained models mainly rely on learning from large amounts of unstructured data. Due to the lack of external knowledge guidance, these models have some problems, such as low learning efficiency, poor model effect and limited knowledge reasoning ability. Therefore, how to use knowledge to enhance the representation ability of pre-trained models is one of the difficulties in the research and application of pre-trained models [22,23]. At present, the mainstream knowledge enhancement pre-training models are mainly divided into two categories. Some models can label the knowledge contained in the text by weak supervision method, and then design a knowledge class pre-training task to learn the knowledge in the text. For example, ERNIE learns from text by tagging and masking phrases and entities in data. In reference [16], entity knowledge was replaced so that the language model could infer entities and relationships in the knowledge graph based on contextual information, thus strengthening the learning of text sequence knowledge. Other models can be co-pre-trained on structured knowledge bases and unstructured texts, such as K-BERT7, CoLAKE, and ERNIE. Through the combined learning of structured knowledge and massive unstructured data, the knowledge-enhanced pre-training model can improve the ability of knowledge memory and reasoning.

(1) Pre-trained model incorporating language knowledge.

Language knowledge is the basis of understanding natural language, which mainly includes lexical knowledge, syntactic structure knowledge and semantic knowledge. There are two methods for the fusion of language knowledge in the pre-trained model: one is to automatically label the language knowledge in the unlabeled text to guide the learning of the pre-trained model; the other is to integrate the artificially constructed language knowledge base. By building a multi-granularity mask language model based on N-Gram, ERNIE-Gram [24] can simultaneously learn the semantic relations within and between N-Grams, enabling the model to capture both fine-grained and coarse-grained language knowledge, and significantly improving the semantic representation capability of the model. In addition to integrating linguistic granular knowledge, there is also work on how to learn semantic relationships in sentences. By modeling reference resolution during pre-training, CorefBERT enhanced the model's ability to learn semantic knowledge [25]. By predicting the hidden and recurring named entities in the text, the algorithm improved the model's ability to model reference relationships.

The above methods mainly annotate the human knowledge contained in the unannotated data, and let the model integrate the language knowledge by learning the annotated information. In addition, there are also studies that integrate artificially constructed language knowledge bases into pre-trained models. Among them, WordNet and HowNet are representative language knowledge bases.

These knowledge bases contain a wealth of language knowledge. WordNet organizes words from different parts of speech into a synonym set. Each synonym set represents a basic semantic concept. WordNet uses semantic relationships to connect these collections into a network. Each word has its own explanation and example. SenseBERT is a combination of WordNet concepts such as supersense. By restoring the obscured words and predicting their corresponding supermeanings, the model can explicitly learn the semantic information of words in a given context. SenseBERT's effectiveness on tasks such as word sense disambiguation has been improved significantly. LIBERT designs the pre-training task process of lexical relation classification by using the synonymy and upper and lower relation between words in WordNet, which enhances the modeling ability of the pre-training model for semantic information and improves the effect on most natural language processing tasks.

(2) Pre-trained models incorporating world knowledge.

In the process of understanding the world, human beings have produced a great deal of world knowledge. Part of the knowledge can be described by entities and the relationships between entities. The researchers represent this world knowledge through knowledge graphs. In a knowledge graph, an entity represents a node in the network, and the relationships between entities represent the edges between the corresponding nodes. It is an important method to store the world knowledge by using the knowledge graph and make the model learn the human cognition of the world explicitly. KEPLER combines the pre-trained context encoder with the knowledge model, so that the pre-trained model can not only better integrate the factual knowledge in the graph triples into the model, but also effectively learn the knowledge representation of entities and relationships through rich entity descriptions. Unlike KEPLER, there are models that combine language and knowledge. CoLAKE regards the text sequence as a fully linked word graph [26], and takes each entity as the anchor point to connect the sub-graphs in the knowledge graph corresponding to the entities in the text to form a word-knowledge graph containing words, entities and relations at the same time. By learning the word-knowledge graph, the model can integrate both the language knowledge in the training corpus and the world knowledge in the graph. However, CoLAKE mainly focuses on the modeling of entities in knowledge graph, but neglects the representation of entities in training corpus. Therefore, ERNIE 3.0 proposes the method of parallel pre-training of knowledge graph and text, using text to express knowledge. ERNIE 3.0 breaks through the bottleneck that heterogeneous structured knowledge representation and unstructured text representation are difficult to model uniformly.

(3) Pre-trained model integrating domain knowledge.

Artificial intelligence industry applications exist a wealth of expertise accumulated by many industry experts. Current pre-training models mainly rely on Internet data for training. The lack of industry-relevant domain knowledge in the data results in poor performance of pre-

trained models on natural language processing tasks in specialized domains. In the medical field, for example, the application of CBLUE has shown that general-purpose pre-trained models are less effective than humans at handling such tasks. In order to enhance the application effect of the pre-training model in the professional field, researchers have explored how to integrate the domain knowledge into the pre-training model. BioBERT is a pre-trained model in the biomedical field [27,28]. Experiments show that pre-training on the biomedical corpus can significantly improve the performance of the model on tasks in the biomedical field. For the pre-training method of domain knowledge, Ernie-Health uses medical entity mask algorithm to learn entity knowledge such as professional terms. At the same time, through the medical question and answer matching task, the model can learn the corresponding relationship between the description of disease conditions and the professional treatment plan of doctors, and obtain the internal relationship between the knowledge of medical entities. The effect of Chinese medical text processing tasks including medical information extraction and medical term normalization has been significantly improved. Furthermore, combining the learning methods of world knowledge and domain knowledge, BERT-MK learns based on the sub-graph of medical knowledge graph, which improves the application effect of the pre-trained model in the medical field tasks.

In order to fully integrate domain knowledge, the models represented by FLAN, ExT5 and T0 collect task data from 60, 107 and 171 domains respectively, and design task templates for each task. By converting a variety of tasks into a unified format from text to text generation, the model can integrate and use multi-domain and multi-task knowledge in the pre-training stage, which can significantly improve the general ability and generalization performance of the model. PPT network continues to transform multiple tasks into a unified format through templates, and can learn domain knowledge of continuous prompts in the pre-training stage, improving the model's ability to transfer fewer samples on downstream tasks where training samples are scarce.

Knowledge enhancement pre-training model can significantly improve its performance by integrating various types of external knowledge. However, in the process of learning knowledge, the model usually has a knowledge forgetting problem, that is, after learning new knowledge, it will forget the previously learned knowledge. Therefore, how to solve the problem of knowledge forgetting is very important. In order to avoid knowledge forgetting, ERNIE 2.0 builds a framework for continuous pre-training. Under this framework, whenever a new task is introduced, the framework can learn that task while still remembering what was previously learned. In addition, K-ADAPTER learns world knowledge and language knowledge through different adapters. In the downstream task, the method can concatenate the feature representation generated by different adapters, and generate the representation with various knowledge at the same time, so that multiple knowledge can be applied to the task at the

same time, and effectively solve the problem of knowledge forgetting.

3. Long text representation model based on knowledge enhancement

Existing models struggle to handle real-world, document-level data, especially when long text contains domain-specific knowledge. In addition, in order to reduce the ambiguity of understanding, these texts are often rigorously structured. For well-structured natural language documents, sentence representation and knowledge should be integrated effectively to better fit the original structure of the text. Therefore, this paper proposes a long text representation model based on knowledge enhancement (KEN). KEN enhances document-level representation based on auxiliary tasks extracted from sentence-level knowledge elements to improve the performance of primary tasks, such as similar case matching. The experiment proves that KEN can integrate the external knowledge and the hierarchical structure of the document into the document level text representation better than the method of directly reducing the document to super long sentences or not adding other relevant information.

3.1. Knowledge enhancement method based on multi-label prediction

In general, the text representation is defined as a mapping $x = F(x)$. Where x represents a piece of text consisting of characters. And x represents a low-dimensional real vector. $F(\cdot)$ is a mapping function. The relationship between document-level data set D with data volume N and background knowledge A with M knowledge elements can be defined as follows:

$$D = \{x_n, y_n\}_{n=1}^N \leftrightarrow A = \{a_m\}_{m=1}^M \quad (1)$$

$$y_n = \{L_1, L_2, \dots\} \leftrightarrow Task = \{Task_1, Task_2, \dots\} \quad (2)$$

Where y_n represents the collection of labels corresponding to document-level data x_n . L represents the unique heat vector of the corresponding task label. a_m represents a knowledge element of background knowledge space A . According to the diversity of text expression, the number and naming of knowledge elements are not fixed. Text features need to be extracted hierarchically, so for any document T in D , there is a feature $Extractor()$, which

can extract the probabilistic representation of each sentence S in the document about A .

$$T = [S_1, \dots, S_j, \dots, S_{|T|}] \quad (3)$$

$$S'_j = Extractor(S_j) \quad (4)$$

$$K_j = f(S_j) = [p_1, \dots, p_m, \dots, p_M] \quad (5)$$

Where p_m represents the conditional probability between knowledge element a_m and residual knowledge element. K_j and S'_j are the knowledge representation and semantic representation of the sentence S_j , respectively. By observing the above formula, it can be seen that knowledge representation K_j can be learned by the model as a classification task.

If there is a support scheme $L_{support}$ in the form of multi-label classification in D , or there is a $L_{support}$ in another data set D' mapped to A , and this task is one application of background knowledge A , then knowledge space A can be replaced by at least one $L_{support}$, that is, when data set D' satisfies formula (6), there is, if \exists is at least one of the application of A ,

$$A = \{a_m\}_1^M \Rightarrow \cup (L_{support} = \{l_h\}_1^H) \quad (6)$$

Because a document is a sequence of sentences, T can be represented as a semantic matrix M_s and a knowledge matrix M_k consisting of S_j and K_j .

$$M_s = [S_1: \dots: S_{|T|}]_{|S| \times |T|} \quad (7)$$

$$M_k = [K_1: \dots: K_{|T|}]_{H \times |T|} \quad (8)$$

Where $H = M$ when $A = \{a_m\}_1^M$ is replaced by exactly one $L_{support}$. For M_s and M_k , they are processed to get v_s and v_k respectively, and then the vector of document T represents T' as a concatenation of two vectors:

$$T = F(T) = [f_s(M_s); f_k(M_k)] \quad (9)$$

When other tasks in y_n are the main tasks that people focus on, T can be used as the input to the main task for calculation.

3.2. Model definition

As shown in Figure 2, the model is divided into three parts: knowledge extraction part, fusion part and application part. The knowledge extraction part and the fusion part jointly generate the document level text representation, and they jointly play the role of vector generation in the model. The application section applies the document vector to the main task, and the application section is specifically designed with the specific form of the main task (classification or regression).

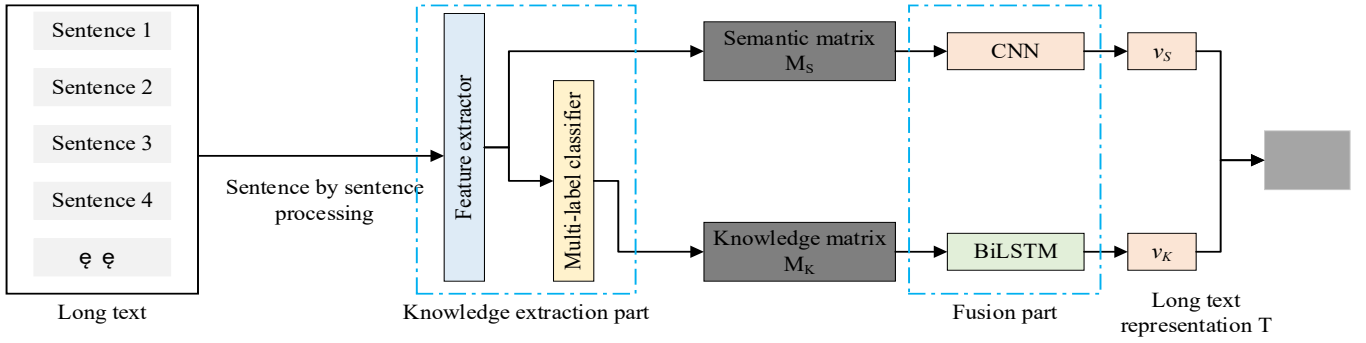


Figure 2. Proposed knowledge enhancement model

Specifically, the knowledge extraction part consists of a feature extractor and a multi-label classifier. Because BERT model based on multi-head attention mechanism has excellent sequence representation ability, $BERT(\cdot)$ is generally used to replace $Extractor()$ in formula (4), and BERT is used to represent the sequence start character "[CLS]" in sequence s . On this basis, formula (4) is changed as:

$$S'_j = BERT(S_j) \quad (10)$$

Since formula (6) is computed and the task of extracting the knowledge elements contained in the text character sequence is obtained, the probability distribution of the category corresponding to the multi-label classification task is obtained, so formula (5) can be generalized as:

$$K_j = \text{sigmoid}(W_K \times S'_j + b_K) \quad (11)$$

Where W_K and b_K represent weight matrix and bias term respectively. It should be pointed out that according to formula (6), when the task corresponding to the original data set contains a multi-label classification task under the same background knowledge, the multi-label classification task can be used as an auxiliary task to participate in training together with the main task in the original data set. When there is no multi-label classification task satisfying the conditions in the original data set, knowledge elements on another data set can be learned by means of the two-stage approach, and then applied to the main task of the original data set. In this paper, the model is trained in two stages.

In order to take into account the dependence between knowledge elements and the problem of category imbalance, the loss function of the multi-label classifier is designated as the focus loss function.

$$y' = [y'_1, \dots, y'_H], y'_n \in \{\mp 1\} \quad (12)$$

$$K'_j = \text{sigmoid}(y' \odot (\alpha K_j + \beta)) \quad (13)$$

$$\text{focal-loss} = \sum_1^H \frac{-\log(K''_j)}{\alpha} \quad (14)$$

Where y' represents the label vector corresponding to the data. The value of each component in the vector is 1 or -1, indicating the existence of the class, and \odot indicates the multiplication of the corresponding positions of the vector. K''_j represents the components of the vector K'_j . α and β

are focal loss parameters, and according to the empirical value [29], let $\alpha=4$ and $\beta=0$.

Each sentence in the original document has two representations of semantics and knowledge, and then the semantic matrix M_s and knowledge matrix M_k are obtained according to the order of the sentence in the document. The function of the information fusion part of the model is to extract the information that fits the document structure from the semantic matrix and the knowledge matrix respectively, and fuse them together. Considering that the information expressed in the semantic matrix is too dense, convolutional neural networks are used to process:

$$\lambda = \text{ReLU}(W_c \times S_{i:i+t} + b_c) \quad (15)$$

$$v_c = \text{maxpooling}([\lambda_1; \lambda_2; \dots]) \quad (16)$$

Where ReLU is the activation function of convolutional kernel in convolutional neural network. t is the size of the sliding window of the convolution kernel during the convolution process, and W_c and b_c are the weight matrix and bias of the convolution kernel when processing the semantic matrix, respectively. λ is the vector obtained by a convolution kernel of the sentence M_s . maxpooling indicates the maximum pooling operation. The length of the vector v_c is the number of convolutional kernel in the convolutional neural network.

$$v_k = \text{BiLSTM}(M_k) \quad (17)$$

$$T = [v_k; v_c] \quad (18)$$

4. Experiments and analysis

The data set involved in the experiment comes from CAIL2019 similar case matching track. Its corresponding task is to match similar cases, that is, to calculate and judge the similarity of several legal documents involving private lending. The data set comprises a total of 15000 legal instruments relating to private lending, each of which provides only a title and factual description [30]. The 15000 documents are packaged into 5000 triples, each consisting of a query document and two candidate documents, and the goal of the task is to find the candidate document most similar to the query document.

In general, an ordinary text matching task is designed to determine whether a pair of text $\langle x_q, x_i \rangle$ is similar. After the text is mapped into the same vector space, the similarity value $score_{q-i}$ is calculated. If the similarity value is greater than or equal to the threshold value τ , a similar conclusion is drawn; otherwise, a dissimilar conclusion is drawn. For similar case matching, it is required to find the document most similar to the inquiry document d_q in the candidate document set $R_q = \{d\}_1^{N'} (N' > 1)$ corresponding to the inquiry document d_q as the result d_r , and the similarity value of any document in R_q is greater than the similarity threshold.

$$\forall d_i \in R_q, score_{q-i} \geq \tau \quad (19)$$

$$d_r = \operatorname{argmax}_{d_i \in R_q} \{score_{q-i}\} \quad (20)$$

Formula (1) can be converted to:

$$D = \{\langle d_{q,i}, R_{q,i} \rangle, \langle d_{r,i} \rangle_1^N\} \leftrightarrow A: \{Knowledge\ elements\ of\ private\ lending\}_1^M \quad (21)$$

The task of multi-label classification is not included in data set D , so data set CAIL2019-FE is introduced as D' . The legal documents published by the "China Judicial Documents Network" are the source of data set D' . Each piece of data in data set D' is a fragment of a case description. It is important to note that each sentence has a different number of category labels, and there are even a large number of sentences with empty labels. The data set covers three areas, including marriage and family, labor disputes, and loan contracts. All data is marked by professionals with a background in legal knowledge. According to the case element system preset by experts in the field, the purpose of case element classification is to classify the description of important facts in the case description according to the system. The results of case element classification can be used in the actual business needs of the judicial field, such as case briefs, interpretable case pushing and relevant knowledge recommendation. Specifically, given a relevant paragraph in a judicial document, the system needs to judge each sentence in the document and identify the key elements of the case.

$A: Private\ lending\ field \Rightarrow$

$L_{support}: Case\ element\ classification\ scheme \quad (22)$

Because of the different granularity of the two data sets, the training of the model is divided into two stages: the auxiliary task stage and the main task stage. In the auxiliary task stage, the case element extraction is the main task, and the feature extractor and multi-label classifier are trained. The fusion part does not participate in the training. The main task stage is based on the matching task of similar cases, and the fusion part and the application part are trained.

In the application part, combined with the specific situation of CAIL-2019-SCM, that is, $N' = 2$ in $R_q = \{d\}_1^{N'}$, the application part is designed as follows on the basis of equation (18).

$$v_{q-i} = \delta(x_q - x_i) \quad (23)$$

$$pre = \frac{1}{N'} \sum_{i=1}^{N'} v_{q-i} \quad (24)$$

$$\hat{y} = \operatorname{softmax}(W_a \times pre + b_a) \quad (25)$$

Where W_a and b_a represent the weight matrix and bias. v_{q-i} represents the relationship between the candidate legal instrument and the inquiry instrument in the candidate set by means of difference vector. δ represents the similarity between x_q and x_i . \hat{y} is a binary prediction result based on pre to determine which of the two candidate documents is most similar to the query document.

In the auxiliary task stage, the BERT-base Chinese version pre-training model released by Google is adopted, and the learning rate during training is 3×10^{-6} . In subsequent experiments, this model was used for all BERT related models. The changes of $F1_{micro}$ and $F1_{mac}$ indicators on the verification set during model training are shown in Figure 3. Their final performance in the test set is $F1_{micro} = 82.88\%$ and $F1_{mac} = 61.96\%$. Therefore, the second stage of the model is based on this model and is specially labeled as $BERT_{trained}$.

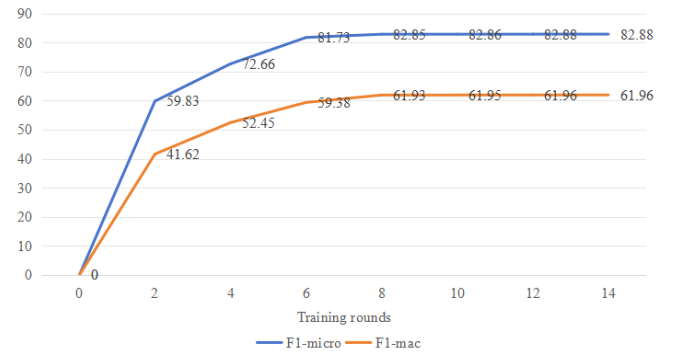


Figure 3. Performance change of BERT model on verification set during auxiliary task phase

In order to compare with the proposed model, two traditional text representation models, two traditional neural network-based text representation models, BERT, two advanced methods are set up for the comparison experiment. In terms of knowledge verification, two text representation models based on $BERT_{trained}$ and integrated with background knowledge information are set as control models.

TF-IDF: Text representation model based on vector space model. This model is commonly used for relevance scoring in text search. The eigenvalue in the text representation is the product of the word frequency of each word in the document and the inverse document frequency.

LDA: Text presentation model based on topic model. The model deduces the topic distribution of the document, gives the topic of each document in the corpus in the form of probability distribution, and analyzes its topic distribution, so as to obtain the vector representation of the document, which can be used for semantic similarity calculation.

CNN: Text representation model based on traditional neural networks. The model uses word embedding based

on neural network training and convolutional neural network to generate text representation, which makes up for the defect of context loss when deep neural network is used alone.

LSTM: A single semantic text matching model based on traditional neural networks. The model uses a long and short term recurrent neural network with memory unit to capture word order information, automatically weakens unimportant words, and takes the output of corresponding neurons at the end of the sequence as the representation of sentences.

BERT: Unlike the language models proposed in recent years, BERT no longer focuses on the information before and after the words, but on the information of the whole context in each layer of the model. With a pre-trained BERT model, it can get good results by simply adding an output layer and fine-tuning the model. By contrast, the model treats the entire document as one super-long sentence. In the experiment, the model was trained with $BERT_{trained}$ in the auxiliary task stage

BERT+KER [31]: This model is based on $BERT_{trained}$ and does not model the entire document hierarchically, but treats the entire document as a single ultra-long sentence. The BERT model can be used to construct the representation of sentence pairs based on appropriate modification of the data. Therefore, the regular expression is summarized from the legal interpretation, and keyword extraction is carried out in the data set, and a new sequence composed of the extracted results is constructed and added to the original document. The model also outputs BERT's representation of the sequence starter.

BERT+K [32]: Based on $BERT_{trained}$, but unlike the BERT+KE model, the knowledge information in BERT+K comes from the multi-label classifier in $BERT_{trained}$. The model also treats the document as a single ultra-long sentence rather than a sequence of sentences.

Combined with the characteristics of the data set, in order to control variables, all the above models except BM25 are matched by formulas (23-25) when establishing the text matching model. In terms of evaluation indexes, accuracy (Acc) and F1 value of positive samples are uniformly adopted as evaluation indexes for the performance of all models on the test set.

$$Acc(f: D) = \frac{1}{m} \sum_{i=1}^m x_i \quad (26)$$

Table 1 is a comparison of experimental results, which generally show that the proposed model in this paper is significantly superior to all other methods. TF-IDF, LDA, CNN and LSTM have successively increased in the ability of text representation, so in general, the dense text vector representation method based on neural network is superior to the relatively sparse traditional text representation method. Compared with other models, BERT model has a considerable improvement, especially in F1 value. This shows that BERT, with the support of multi-head self-attention mechanism, can dynamically show the changes of character-level representation in text, so it can get better text representation from character-level representation. It can be seen from the table that knowledge is an important

factor to improve the performance of the model. In the BERT+KER model, regular expressions derived from legal interpretations are actually another form of introducing knowledge. With the introduction of knowledge, words that are originally cut into words are reassembled into words with practical meaning and encoded. Compared with BERT+KER model, the accuracy of BERT+K is greatly improved. In the comparison between BERT+KER and proposed model, proposed model has obvious advantages. For a more intuitive representation, we present the results as shown in Figure 4.

Table 1. Experiment result

Model	Acc/%	F1/%
TF-IDF	55.3	51.7
LDA	59.1	57.2
CNN	62.4	65.8
LSTM	60.7	64.2
BERT	63.8	67.9
BERT+KER	73.8	76.3
BERT+K	77.5	70.0
Proposed	84.7	82.6

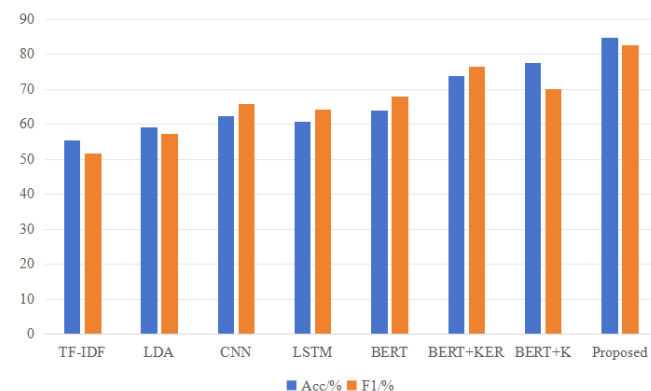


Figure 4. Visualization result

There are two possible reasons for the outstanding performance of the proposed model. Firstly, BERT, as a sentence feature extractor, is very good at capturing the word level information inside the short text (sentence) to generate sentence level representation. Secondly, the model adopts the way of extracting information sentence by sentence, which not only considers the micro-word level information, but also grasps the document-level information from the macro-point of view. With the help of knowledge extraction, the model can not only identify the key case elements hidden in the document statements, but also capture the subtle differences within the elements, which enhances the model's ability to recognize similar documents.

The ablation experiment is shown in Table 2. Formula (4) is replaced by bidirectional convolutional neural network (Bi-CNN) and Bidirectional gated recurrent neural network (Bi-GRU), respectively. In addition, three models are set up on the basis of Bi-CNN: the Bi-CNN which only outputs the representation of legal documents; the Bi-CNN+K which replaces BERT in the BERT+K model with the Bi-CNN model; and the Bi-CNN+H which outputs text without knowledge representation sentence by sentence. It should be noted in particular that the length of the sequence that can be captured by the bidirectional convolutional neural network is set to 512 due to the consideration of control variables.

Table 2. Ablation results

Model	Acc/%	F1/%
Bi-CNN+proposed	72.3	73.2
Bi-GRU+proposed	74.1	73.4
BERT+proposed	84.4	82.2
Bi-CNN	59.8	61.9
Bi-CNN+K	63.8	69.3
Bi-CNN+H	67.7	61.9

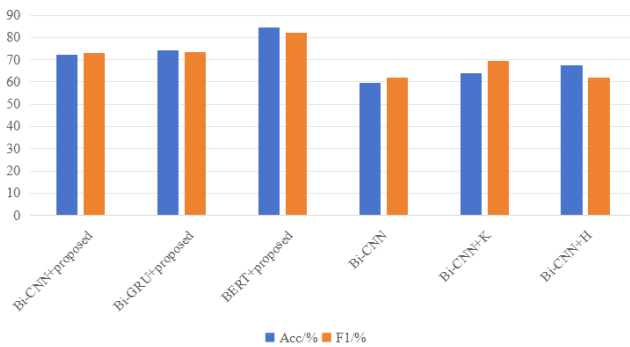


Figure 5. Visualization result for ablation results

By comparing the experimental results of Bi-CNN+proposed, Bi-GRU+proposed and BERT+proposed in Table 2, It can be found that when the length of the sequence extracted by the feature extractor is the same, the performance of the feature extractor has a great impact on the model. From another point of view, it is reasonable to choose BERT as the feature extractor. After obtaining the representation of the sentence, the model not only needs to reuse the representation of the sentence to extract the knowledge element, but also needs to combine the knowledge element with the representation of the corresponding sentence to obtain a richer semantic representation. Therefore, the feature extractor in the knowledge extraction part of the model is a very important component. In the case of the same feature extractor, if the text information contains both the character information of the text itself and the information of the corresponding knowledge concept in the text, it will have a considerable impact on the improvement of the model.

5. Conclusion

Aiming at the problem that long text data with specialized domain knowledge is difficult to understand, this paper proposes a long text representation model based on knowledge enhancement. By establishing the relationship between multi-label classification space and knowledge representation space, a knowledge enhancement method based on multi-label prediction is proposed. The domain-related multi-label classification prediction is taken as the representation of domain knowledge, and it is integrated into the corresponding text representation, so as to integrate the text-related domain knowledge information. In view of the hierarchical structure of long text data, a classical deep neural network is used to process the basic semantic information and knowledge information respectively, and finally fuse into a new knowledge enhanced long text representation. The comparison and ablation results of similar case matching tasks in the field of law not only prove the effectiveness of the proposed method, but also illustrate the importance of adding text structure information and background knowledge information to the text representation model.

Acknowledgments

This work was supported by the Special Fund of Basic scientific Research Business expenses of undergraduate universities in Liaoning Province. Project name: Application of a large language model for enhancing career ability map driven by knowledge base in education and teaching scenarios. Project number: LJ232410166062.

References

- [1] Gao P, Li J, Liu S. An introduction to key technology in artificial intelligence and big data driven e-learning and e-education[J]. Mobile Networks and Applications, 2021, 26(5): 2123-2126.
- [2] Zha D, Bhat Z P, Lai K H, et al. Data-centric artificial intelligence: A survey[J]. ACM Computing Surveys, 2025, 57(5): 1-42.
- [3] Ahmad K, Iqbal W, El-Hassan A, et al. Data-driven artificial intelligence in education: A comprehensive review[J]. IEEE Transactions on Learning Technologies, 2023, 17: 12-31.
- [4] Yu J, Lu Z, Yin S, et al. News recommendation model based on encoder graph neural network and bat optimization in online social multimedia art education[J]. Computer Science and Information Systems, 2024, 21(3): 989-1012.
- [5] Liu Y, Xu Y, Zhou S. Enhancing User Experience through Machine Learning-Based Personalized Recommendation Systems: Behavior Data-Driven UI Design[J]. Authorea Preprints, 2024.
- [6] Yan F, Zhang X, Yang C, et al. Data-driven modelling methods in sintering process: Current research status and perspectives[J]. The Canadian Journal of Chemical Engineering, 2023, 101(8): 4506-4522.

- [7] Wang Z, Wang Y. Digital Library Book Recommendation System Based on Tag Mining[J]. Journal of Artificial Intelligence Research, 2024, 1(1): 10-16.
- [8] Wang H, Li B, Gong J, et al. Machine learning-based fatigue life prediction of metal materials: Perspectives of physics-informed and data-driven hybrid methods[J]. Engineering Fracture Mechanics, 2023, 284: 109242.
- [9] Wang H, Li J, Wu H, et al. Pre-trained language models and their applications[J]. Engineering, 2023, 25: 51-65.
- [10] Ding N, Qin Y, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.
- [11] Zhang X, Malkov Y, Florez O, et al. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter[C]//Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. 2023: 5597-5607.
- [12] Zhu R J, Zhao Q, Li G, et al. Spikept: Generative pre-trained language model with spiking neural networks[J]. arXiv preprint arXiv:2302.13939, 2023.
- [13] Arisoy E, Chen S F, Ramabhadran B, et al. Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2013, 22(1): 184-192.
- [14] Choudhary K, Beniwal R. Xplore word embedding using CBOW model and skip-gram model[C]//2021 7th international conference on signal processing and communication (ICSC). IEEE, 2021: 267-270.
- [15] Xiong Z, Shen Q, Xiong Y, et al. New Generation Model of Word Vector Representation Based on CBOW or Skip-Gram[J]. Computers, Materials & Continua, 2019, 60(1).
- [16] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
- [17] Yao T, Zhai Z, Gao B. Text classification model based on fasttext[C]//2020 IEEE International conference on artificial intelligence and information systems (ICAIS). IEEE, 2020: 154-157.
- [18] Ma L, Yang W, Xu B, et al. Knowlog: Knowledge enhanced pre-trained language model for log understanding[C]//Proceedings of the 46th IEEE/ACM international conference on software engineering. 2024: 1-13.
- [19] Kamaloo E, Clarke C L A, Rafiei D. Limitations of Open-Domain Question Answering Benchmarks for Document-level Reasoning[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 2123-2128.
- [20] A. Karak, K. Kunal, N. Darapaneni, and A. R. Paduri. Implementation of GPT models for Text Generation in Healthcare Domain[J]. EAI Endorsed Trans AI Robotics, vol. 3, Apr. 2024.
- [21] S. Fang. A Survey of Data-Driven 2D Diffusion Models for Generating Images from Text[J]. EAI Endorsed Trans AI Robotics, vol. 3, Apr. 2024.
- [22] Li B, Jiang G, Li N, et al. Research on large-scale structured and unstructured data processing based on large language model[C]//Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering. 2024: 111-116.
- [23] Li I, Pan J, Goldwasser J, et al. Neural natural language processing for unstructured data in electronic health records: a review[J]. Computer Science Review, 2022, 46: 100511.
- [24] Hu L, Liu Z, Zhao Z, et al. A survey of knowledge enhanced pre-trained language models[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(4): 1413-1430.
- [25] Zhu H, Peng H, Lyu Z, et al. Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation[J]. Expert Systems with Applications, 2023, 215: 119369.
- [26] Sun T, Shao Y, Qiu X, et al. Colake: Contextualized language and knowledge embedding[J]. arXiv preprint arXiv:2010.00309, 2020.
- [27] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
- [28] Bhasuran B. BioBERT and similar approaches for relation extraction[M]//Biomedical Text Mining. New York, NY: Springer US, 2022: 221-235.
- [29] Zhao Y, Li H, Yin S. A Multi-channel Character Relationship Classification Model Based on Attention Mechanism[J]. Int. J. Math. Sci. Comput.(IJMSC), 2022, 8: 28-36.
- [30] M. Tyagi, P. K. Singh, S. K. Yadav, and S. K. Soni. A Multi-Channel Spam Detection System Utilizing Natural Language Processing and Machine Learning[J]. EAI Endorsed Trans AI Robotics, vol. 4, Mar. 2025.
- [31] Wang H, Li J, Li Z. AI-generated text detection and classification based on BERT deep learning algorithm[J]. arXiv preprint arXiv:2405.16422, 2024.
- [32] Yu B, Tang F, Ergu D, et al. Efficient classification of malicious urls: M-bert—a modified bert variant for enhanced semantic understanding[J]. IEEE Access, 2024, 12: 13453-13468.