information

MDPI

Article

Medical QA Oriented Multi-Task Learning Model for Question Intent Classification and Named Entity Recognition

Turdi Tohti , Mamatjan Abdurxit and Askar Hamdulla *

School of Information Science and Engineering, Xinjiang Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi 830017, China

* Correspondence: askar@xju.edu.cn; Tel.: +86-139-9922-1222

Abstract: Intent classification and named entity recognition of medical questions are two key subtasks of the natural language understanding module in the question answering system. Most existing methods usually treat medical queries intent classification and named entity recognition as two separate tasks, ignoring the close relationship between the two tasks. In order to optimize the effect of medical queries intent classification and named entity recognition tasks, a multi-task learning model based on ALBERT-BILSTM is proposed for intent classification and named entity recognition of Chinese online medical questions. The multi-task learning model in this paper makes use of encoder parameter sharing, which enables the model's underlying network to take into account both named entity recognition and intent classification features. The model learns the shared information between the two tasks while maintaining its unique characteristics during the decoding phase. The ALBERT pre-training language model is used to obtain word vectors containing semantic information and the bidirectional LSTM network is used for training. A comparative experiment of different models was conducted on Chinese medical questions dataset. Experimental results show that the proposed multi-task learning method outperforms the benchmark method in terms of precision, recall and F_1 value. Compared with the single-task model, the generalization ability of the model has been improved.

Keywords: multi-task learning; named entity recognition; intent classification; ALBERT; deep learning

check for updates

Citation: Tohti, T.; Abdurxit, M.; Hamdulla, A. Medical QA Oriented Multi-Task Learning Model for Question Intent Classification and Named Entity Recognition. *Information* **2022**, *13*, 581. https://doi.org/10.3390/info13120581

Academic Editor: Krzysztof Ejsmont

Received: 30 October 2022
Accepted: 10 December 2022
Published: 14 December 2022

1. Introduction

Along with the rapid development of online medical service technology, people can ask questions and receive answers online through health service websites. Medical question and answer systems require the ability to build a medical knowledge base and apply natural language understanding techniques to extract structured information from users' questions, and automatically generate answers to them. For this type of health service, the accuracy of the generated answers depends not only on the quality of the knowledge base, but also on the accuracy of the user's question understanding.

Intent classification and named entity recognition are two subtasks of natural language understanding. Most existing medical natural language processing studies in text classification and named entity recognition are usually performed independently. The purpose of the intent classification task is to first identify possible intent classes in a given domain and then classify sentences to specific intent classes based on contextual information in the text. Named entity recognition aims at extracting medical entities from the text and predicting the different kinds of entities. Both of these tasks can help a medical question and answer system to correctly provide the services required by the user. For example, suppose a user asks the question "I have kidney stones, what should I do?". Based on the intent analysis, the user is seeking a treatment and based on named entity recognition, we know that the question contains the disease term "kidney stones". In this case, we can search our knowledge base and return an answer about the treatment for kidney stones. In some tasks where the amount of data is small, training with multi-task learning allows the network to rely on tasks where data is easily available to learn how to extract the underlying information or learn to extract some features common to both tasks, and to obtain better generalization capabilities. Most current medical questions intent classification and named entity recognition methods use single-task learning strategies that ignore the close relationship between these two tasks. To address this problem, this paper proposes a multi-task learning model combining ALBERT pre-training model and BILSTM, where named entity recognition task and intent classification task share ALBERT word embedding layer and BILSTM layer. They facilitate each other's learning to obtain rich semantic and associative information at word and sentence level by sharing the underlying parameters, Experiments were conducted on the Chinese healthcare questions dataset. In summary, our main contributions can be summarized as follows:

- A multi-task learning model based on ALBERT-BILSTM is proposed for intent classification and named entity recognition of Chinese online medical questions.
- The experimental results demonstrate that the proposed method in this paper outperforms the benchmark methods and improves the model generalization ability compared to the single-task model.

2. Related Work

2.1. Medical Named Entity Recognition

Medical named entity recognition is used to recognize blocks of words in a text that are related to specific entities, such as symptoms, drugs, and treatments. Rule-based approaches play an important role in named entity recognition. Gerner et al. [1] used a dictionary-based approach to identify species names in biomedical literature. Fukuda et al. [2] proposed a rule-based approach to extract names of substances such as proteins from biological documents. However, these methods require manual rules and, thus, lack generality. Researchers have also attempted to recognize entities from unstructured data using machine learning methods. He et al. [3] proposed a conditional random field (CRF) based approach to recognize drug names in biomedical texts. Machine learning methods rely on manual feature design, which is both time consuming and laborious. In recent years, deep learning methods that can improve the performance of named entity recognition without relying on feature engineering have received increasing attention. For example, Chen et al. [4] used a Bi-directional Long Short-Term Memory (BILSTM) model for named entity recognition of adverse drug event reports in China. The BILSTM-CRF based model was validated by many works as a very effective method to solve named entity recognition [5–7]. Current research on named entity recognition of medical text is mainly distributed in medical literature, electronic medical records and less on medical question and answer text. In recent years, some researchers have also started to focus on the research on medical question and answer text. Su et al. [8] used CRF to conduct named entity recognition research on their own constructed dataset, and the extracted entities included diseases, drugs, symptoms, treatments and tests. Qin et al. [9] proposed a BERT-BiGRU-CRF neural network model to recognize named entities in electronic medical records of cerebrovascular diseases in order to address the issues associated with neglecting context information in electronic CVD medical entity recognition. In order to win the CHIP2018 competition, Ji et al. [10] proposed a cooperation approach based on multiple neural network models for Chinese medical named entity recognition.

2.2. Intent Classification

Deep learning methods are widely used for text classification [11]. Ravuri et al. [12] proposed applying LSTM models to the intent classification problem. Zhang et al. [13] used convolutional neural networks (CNN) to analyze online cancer community discussion topics and the CNN outperformed support vector machine (SVM) models and LDA topic models. In addition, researchers have used other knowledge and rule-based functions to improve the classification accuracy of CNNs on clinical texts [14]. Jang et al. [15] proposed an attention-based Bi-LSTM+CNN hybrid model that capitalize on the advantages of LSTM and CNN with an additional attention mechanism. Deep learning methods are still plagued by insufficient training data. Zhang et al. [16] proposed a Capsule network model for electronic medical record classification. Recently, pre-trained models generating representations of words with a priori semantic knowledge in large-scale unlabeled corpora have achieved state-of-the-art results in various natural language processing tasks [17]. Devlin et al. [18] published in 2018 the BERT (Bidirectional Encoder Representations from Transformers) pre-trained language model. This model efficiently classifies TCM records and can obtain the best results by training the corpus with a bidirectional Transformer encoder to obtain text representations [19,20]. Various pre-training models have emerged after BERT. Lan et al. [21] proposed the ALBERT (A Lite BERT) model, which is a lightweight pre-trained language model based on the BERT model and use a bidirectional Transformer to obtain feature representations of the text, but ALBERT greatly reduces the parameters in the model and achieves the best results in several NLP tasks. Zhang et al. [22] presented a short text classification algorithm for Chinese clinical medicine combining ALBERT pre-training model and graph attention network.

2.3. Multi-Task Learning

Multi-task Learning (MTL) refers to learning multiple tasks simultaneously, and neural networks generally allow the underlying network to simultaneously take into account different tasks by sharing weights to extract more representative low-level features, because the purpose of single-task learning often focuses only on local information, limiting the generalizability of the model. In contrast, multi-task learning can exploit the potential information between tasks to extract features that are common across tasks and improve model performance [23]. Two common models of multitask learning exist in deep learning: parameter hard sharing mechanisms and parameter soft sharing mechanisms [24]. The parameter hard sharing mechanism usually works by sharing the hidden layer among all tasks and keeping the task-specific output layer. However, under the parameter soft sharing mechanism, each task has its own model parameters, and the similarity of model parameters is ensured by regularizing the parameter distances. Deep neural network-based multitask learning in natural language processing has been shown to be effective [25,26], especially in the presence of insufficient training data. Multi-task learning strategies have been used to improve the performance of named entity recognition for medical text. Researchers [27] used a multi-task bidirectional long short-term memory network (BILSTM)-based model for the named entity recognition task and lexical tagging task to improve named entity recognition in Chinese electronic medical records. Researchers also proposed a multitask learning framework for named entity recognition (NER) and named entity normalization (NEN) [28], which greatly improved the performance of NER and NEN. These studies used a multi-task learning model to solve two-sequence tagging problems. Peng et al. [29] presented a multi-task learning model with multiple decoders on varieties of biomedical and clinical natural language processing tasks such as named entity recognition, relation extraction, text similarity and text inference.

3. Methodology and Model

The overall structure of the proposed ALBERT-BILSTM-based multi-task learning model is shown in Figure 1, which mainly consists of an ALBERT word embedding layer, a BILSTM network layer and a task-specific decoding layer. The model adopts a hard sharing model of parameters, where the named entity recognition task and the intent classification task share the ALBERT word embedding layer and the BILSTM layer. During the training period, the model alternates between intention classification and named entity recognition tasks. The model directly uses online medical question text as input, and first the word embedding layer converts the input sequence into a corresponding word vector sequence using the ALBERT pre-training model. The word vector sequence is input to the BILSTM layer for further semantic encoding using the forward and backward networks to obtain the final representation of the utterance. Finally, for the intent classification task, the hidden states generated by the BILSTM layer are fed to the fully-connected and softmax classifier to obtain the probabilities of the input text in each intent category, thus achieving the final intent classification. For the named entity recognition task, CRF obtains the best sequence labels for the full sentences by decoding word labels with semantic features. The gradient information from both the intent recognition and named entity recognition tasks is passed backwards to the shared encoder part of ALBERT and BILSTM and the model parameters are updated to obtain a more representative underlying representation.

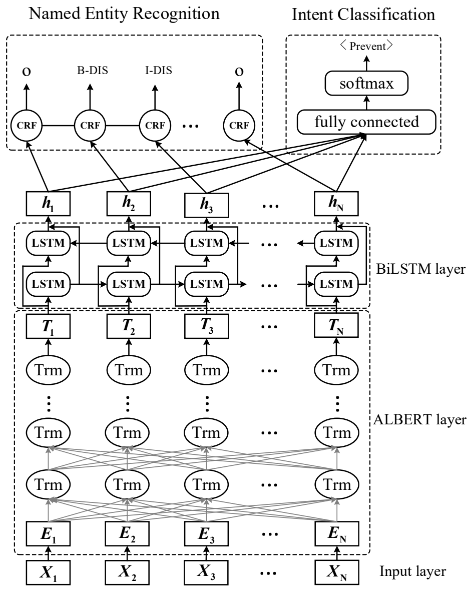

Figure 1. The architecture of MTL-ALBERT-BILSTM model.

3.1. ALBERT Pre-Trained Language Model

ALBERT pre-trained language model is based on a large corpus and uses unsupervised learning methods to learn feature representations for words, which can characterize the polysemy of words and enhance the semantic representation of sentences. ALBERT is a lightweight language model based on the BERT model. To reduce the parameters of the BERT pre-trained model and enhance semantic understanding, the ALBERT model effectively reduces the BERT model parameters by two methods: parameter sharing between layers and parameter factorization of the embedded layers, which can reduce the memory overhead during training and improve the training speed of the model. To compensate for the shortcomings of the NSP task in BERT proposed by Yang et al. [30], ALBERT utilizes SOP (Sentence Order Prediction) instead of NSP (Next Sentence Prediction) task, through which the SOP task model can learn more semantic relations between sentences.

ALBERT model uses bidirectional Transformer encoder (Trm) to obtain the feature representation of text E_1, $E_{2,}$ … E_N, which represents each character in the sequence, and after the training of multi-layer bidirectional Transformer encoder, finally we obtain the text feature vector T_1, T_2, … T_N.

Transformer is an Encoder-Decoder [31] structure based on the Self-Attention mechanism. ALBERT is composed by stacking multiple encoder layers, and the model structure of this part is shown in Figure 2. Each encoder layer includes a Self-Attention layer and a feedforward neural network, and with the help of the Self-Attention mechanism enables the model to allow the current node to not only focus on the current word, but to perform relational computation from the global view to obtain the semantics of the context. In addition, to address the degradation problem in deep learning, the Transformer encoder unit contains an Add&Norm layer for each subnetwork layer, which adds and normalizes the input and output of this layer [32] and uses a residual connection between two subnetwork layers [33].

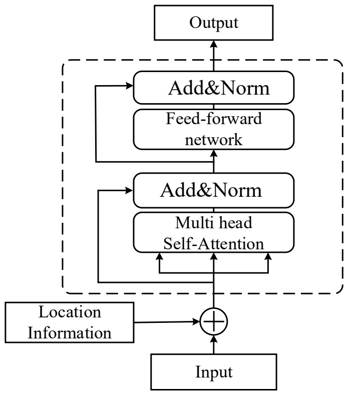

Figure 2. Transformer encoder model architecture.

The most important module of the Transformer encoder is the multi-headed Self-Attention mechanism, which is computed as follows.

$$\begin{gathered}\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \\ \text{Concate}(head_1, head_2, \ldots, head_h)\boldsymbol{W}^O\end{gathered} \quad (1)$$

$$head_i = \text{Attention}(\boldsymbol{QW}_i^{\boldsymbol{Q}}, \boldsymbol{KW}_i^{\boldsymbol{K}}, \boldsymbol{VW}_i^{\boldsymbol{V}}) \quad (2)$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (3)$$

where $\boldsymbol{W}^{\mathbf{o}}$ is the matrix of additional weights, which can compress the dimension of the spliced matrix to the length of the sequence. $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ denotes the query, key and value vectors of each word in the input sequence and W_i^Q, W_i^K, W_i^V are the weights of $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ respectively matrix. $\boldsymbol{d_k}$ denotes the dimension of query and key vectors of each word.

3.2. BILSTM Module

Long short-term memory network (LSTM) [34] is a special type of recurrent neural network that captures the contextual order information of sequences to solve the long dependency problem. LSTM is a variant of RNN that introduces some gate structures to solve the RNN gradient explosion and gradient disappearance problems. The LSTM cell state structure is shown in Figure 3.

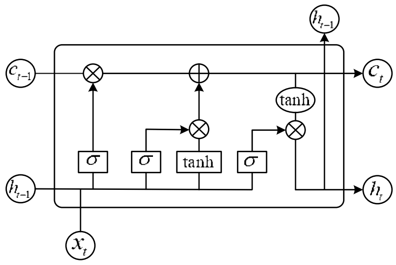

Figure 3. LSTM cell state structure.

LSTM consists of four main components: the storage cell $\boldsymbol{c_t}$, the input gate $\boldsymbol{i_t}$, the output gate $\boldsymbol{o_t}$, and the forget gate $\boldsymbol{f_t}$. In the LSTM, a gate is a method of selectively passing information through, using specially designed "gates" to introduce or remove information from the cell state $\boldsymbol{c_t}$. The LSTM computes an output vector based on the current input and the output of the previous cell, which is then used as the input to the next cell. The calculation formula is as follows.

$$i_t = \sigma(x_t \cdot w_{xh}^i + h_{t-1} \cdot w_{hh'}^i + b_h^i) \quad (4)$$

$$f_t = \sigma(x_t \cdot w_{xh}^{f} + h_{t-1} \cdot w_{hh'}^{f} + b_h^{f}) \quad (5)$$

$$o_t = \sigma(x_t \cdot w_{xh}^{o} + h_{t-1} \cdot w_{hh'}^{o} + b_h^{o}) \quad (6)$$

$$\widetilde{c}_t = \tanh(x_t \cdot w_{xh}^{c} + h_{t-1} \cdot w_{hh'}^{c} + b_h^{c}) \quad (7)$$

$$c_t = i_t \otimes \widetilde{c}_t + f_t \otimes c_{t-1} \quad (8)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (9)$$

where σ denotes the sigmoid activation function, $\boldsymbol{x_t}$ is the cell input at time t and $\boldsymbol{w}$, $\boldsymbol{b}$ denotes the weight matrix and bias vector of the input gate, forget gate, and output gate. $\widetilde{c}_t$ is the intermediate state obtained from the current input. tanh is the hyperbolic tangent function. $\boldsymbol{c_t}$ represents the state at moment t, and $\boldsymbol{h_t}$ is the output at moment t.

LSTM can only encode historical information and ignore future contextual information. In this paper, a bidirectional LSTM (BILSTM) network consisting of forward LSTM and inverse LSTM is used. BILSTM obtains the final hidden layer representation by splicing two different hidden layer representations obtained by sequential and inverse order computation.

3.3. Decoding Unit

(1) Fully connected and normalized exponential function

The fully connected layer reduces the dimensionality of the output vector of BILSTM to the same dimensionality as the total number of medical entity labels. Assuming that S represents the total number of labels, the output of the fully connected layer at the i th position is shown in Equation (10).

$$\boldsymbol{m}_i = \sigma(\boldsymbol{W}_m \boldsymbol{h}_i + \boldsymbol{b}_m) \quad (10)$$

where $W_m \in R^{s*2h}, b_m \in R^s$.

In multi-classification problems, a normalized exponential function (Softmax function) is usually used as the activation function of the output layer of the network. Softmax function can perform a normalization operation on the output values, transforming all output values into probability values between (0, 1), and all probability values add up to 1. This probability value represents the probability that a word belongs to a certain label. The prediction result selects the tag with the highest probability value.

$$\hat{y}_{i,j} = \frac{e^{c_{i,j}}}{\sum\limits_{j=1}^{s} e^{c_{i,j}}} \quad (11)$$

In the training process, the optimal model parameters are learned by minimizing the cross-entropy loss function strategy. The loss function is shown in Equation (12), so that $\boldsymbol{y}_{i,j}$ denotes the true probability that the ith word belongs to the jth label, and takes the value of 0 or 1. $\hat{y}_{i,j}$ is the model prediction value derived from Equation (11).

$$loss = \sum_{i=1}^{n} \left\{ -\sum_{j=1}^{s} y_{i,j} \, log\hat{y}_{i,j} \right\} \quad (12)$$

(2) Conditional random field

There are interdependencies between the labels of named entities of medical texts. For example, the next label of the label “I-disease” will not be “I-drug”. It is a widespread practice to use conditional random field (CRF) optimization to predict the sequence of labels, where the CRF layer takes the sequence $x = (x_1, x_2, \cdots, x_n)$ as input and predicts the most likely sequence of labels $y = (y_1, y_2, \cdots, y_n)$. Given a training set D, all CRF layer parameters (denoted as θ) are estimated by maximizing log-likelihood as shown in Equation (13).

$$L(\theta) = \sum_{(s,y)\in D} \log p(y|x,\theta) \tag{13}$$

where $\boldsymbol{y}$ is the corresponding label sequence of sentence $\boldsymbol{s}$ and $\boldsymbol{p}$ is the conditional probability of $\boldsymbol{y}$ given $\boldsymbol{s}$ and $\boldsymbol{\theta}$. Assuming that $S_\theta(x,y)$ is the score of the sentence label sequence y, the conditional probability p can be calculated using the normalization of $S_\theta(x,y)$. To take advantage of the dependencies between adjacent labels, the model combines the transfer probability matrix $\boldsymbol{T}$ and the emission probability matrix $\boldsymbol{E}$ to calculate the score of the label sequence $S_\theta(x,y)$, as shown in Equation (14).

$$S_\theta(x,y) = \sum_{t=1}^{n} (E_{y_t,t} + T_{y_{t-1},y_t}) \tag{14}$$

where $E_{y_t,t}$ is the probability of word $\boldsymbol{x_t}$ with label $\boldsymbol{y_t}$ and T_{y_{t-1},y_t} is the probability of word $\boldsymbol{x_{t-1}}$ with label $\boldsymbol{y_{t-1}}$ followed by word $\boldsymbol{x_t}$ with label $\boldsymbol{y_t}$. We can find the best sequence of labels for the input sentences by maximizing the log-likelihood over all training sets $\boldsymbol{D}$ by dynamic programming and maximizing the score by using the Viterbi algorithm.

3.4. Multi-Task Learning Step

In the multi-task learning process, both the intent classification and named entity recognition tasks have their own separate training sets $\boldsymbol{D_C}$ and $\boldsymbol{D_N}$. In order to allow both tasks to learn simultaneously, the model alternates between the intent classification and named entity recognition tasks during the training period to “approximate” simultaneous learning. The loss and optimization functions for the entity recognition and intent classification tasks are independent during alternate learning. The intent classification task uses the standard cross-entropy loss as the loss function, and multi-task learning is performed by alternating calls to each task optimizer with the two tasks using the Adam optimization function to learn the parameters of the multi-task model. This means that we can continuously transfer some information from each task to the other task, which is achieved through a shared layer. At each iteration, a task is randomly selected and then some random training samples are chosen from this task to compute the gradient and update the parameters. The exact procedure of the alternating training phase in multitask learning is shown in Algorithm 1.

Algorithm 1 Multi-task learning training process

```
Input: two task datasets $D_C$ and $D_N$
    Batch size $K$ for each task
    Maximum number of iterations $T$, learning rates $\alpha$ and $\beta$
Random initialization parameter $\theta_0$
for $t$ = 1 ··· $T$ do
    /*Prepare the data for both tasks*/
    Randomly divide $D_C$ and $D_N$ into small batch sets
    BC = {$J_{C,1}, \cdots, J_{C,n}$}
    BN = {$J_{N,1}, \cdots, J_{N,m}$}
    end
    Merge all small batch samples $B' = B_C \cup B_N$
    Random sorting $B'$
    for each $J \in B'$ do
        Calculate the loss $L(\theta)$ on the small batch sample
         /* calculate only the loss of J on the corresponding task */
        Update the parameters: $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \nabla_\theta L(\theta)$
    end
end
```

4. Experiments and Results Analysis

4.1. Dataset

The experimental data used in this paper were derived from the Chinese medical question and answer dataset (cMedQA2) [35] and the Chinese medical question sentence intent classification dataset (CMID) [36]. To ensure that each question includes only a single intention, we eliminate samples with unclear intentions and multiple intentions, and select 10,496 user questions from them for the experiments, using the following seven types of intentions to label these questions: definition, prevention, symptom, cause, treatment method, indication and complication. The statistical information of each label is listed in Table 1.

Table 1. Statistics of sample number of intent query.

Category	Quantity
Definition	963
Symptom	1215
Causes	2177
Prevention	1205
Treatment	2506
Indications	1572
Complications	879

The named entity recognition dataset was also derived from questions used for intent classification labeling, for a total of 5856 medical questions. The named entity recognition task uses three types of entities to label interrogative sentences: disease, symptom, and medicine. Chinese medical questions were not subsumed due to the variability of online health texts, sometimes with incorrect expressions, for which current Chinese subsumption tools are unable to meet the demand. Using the BIO annotation model, the interrogative text is manually annotated with the format B-X, I-X and O. B denotes the beginning part of the entity, I represents the Chinese character inside the entity, and O denotes the non-entity. x represents the category of the named entity, which are Disease, Symptom and Drug, representing disease, symptom and drug, respectively. The task is labeled with a total of 7 tags. After obtaining the intent classification and named entity recognition dataset with annotations, we selected 800 user questions as the test set for the intent classification and the named entity recognition task.

4.2. Evaluation Metrics

To evaluate the effectiveness of the model, Precision P, Recall R and F_1 are used to evaluate the effectiveness of the model. The precision rate P is the proportion of correctly predicted samples among all samples with positive predictions; the recall rate R is the proportion of correctly predicted samples among all samples with true positive predictions; and F_1 is the summed mean of the precision rate and recall rate. The calculation formula is as follows.

$$P = TP/(TP + FP) \tag{15}$$

$$R = TP/(TP + FN) \tag{16}$$

$$F_1 = 2PR/(P + R) \tag{17}$$

where *TP* is the true case: positive samples predicted as positive by the model; *FP* indicates the false positive case: negative samples predicted as positive by the model; *TN* is the true negative case: negative samples predicted as negative by the model; *FN* represents the false negative case: positive samples predicted as negative by the model.

4.3. Experimental Environment and Parameter Settings

The experiments in this paper are based on Python 3.7 and implemented using the Pytorch deep learning framework. The CPU is Intel(R) Xeon(R) E5-2678 v3 @ 2.50 GHz. GPU graphics card is GeForce RTX 3090 and the running memory is 24 G.

The parameters of ALBERT-BILSTM multitask model mainly include the parameters in ALBERT and BILSTM, and the values of the variable parameters are changed sequentially to obtain the optimal parameters of the model while fixing the other parameters. In this experiment, the Chinese pre-training model "ALBERT-Base" released by Google is used for ALBERT-Base, which has 12 layers. The hidden layer is 768 dimensions and the 12-head model is used with 110M parameters in total. The number of nodes in the hidden layer is 128 and the number of layers in the BILSTM model is 1. Activation function of the model is ReLU and the ratio of Dropout is set to 0.1 in the training phase. Batch size of the ALBERT-BILSTM multi-task model is set to 32 and the number of iterations is set to 30. The learning rate of the intention classification task is set to 0.001 and the learning rate of the named entity recognition task is set to 0.002. Adam (Adaptive Moment Estimation) optimization algorithm is used to learn the parameters of the multi-task model.

4.4. Benchmark

To verify the effectiveness of the proposed ALBERT-BILSTM-based multitask learning model for the medical questions intent classification and named entity recognition, several models with good results on text classification and named entity recognition were selected as comparisons.

In the comparison experiments, the TextCNN, BILSTM and BILSTM-CRF models are combined with word2vec word granularity word vectors and the word2vec parameters are set as follows: the dimension of the word vector is set to 200, the window size is 5 and the training epoch is 20.

4.5. Experimental Results and Analysis

We trained our models using the named entity recognition and the intent classification training dataset and tested the performance on the named entity recognition and the intent classification task, respectively.

For the named entity recognition task, the model performance was evaluated using the micro-averaged F_1 values, recall and accuracy. We chose BILSTM [4] and BILSTM-CRF [5] as benchmark methods. We also compared with the ALBERT model. In addition, to evaluate whether a multi-task learning strategy can improve the performance, we tested the performance of the model using a single-task learning model ALBERT-BILSTM-CRF. The experimental results are shown in Table 2.

Table 2. Results of NER on different models.

Method	P	R	F_1
BILSTM	0.7515	0.7432	0.7473
BILSTM-CRF	0.7578	0.7641	0.7606
ALBERT-CRF	0.7869	0.7954	0.7911
ALBERT-BILSTM-CRF	0.7926	0.8014	0.7970
MTL-ALBERT-BILSTM	**0.8103**	**0.8036**	**0.8069**

From the experimental results in Table 2, we can find that our method outperforms the baseline method in terms of F_1 value, precision and recall metrics. We found that the BILSTM-CRF model has a higher F_1 value compared with the BLS TM model and the BILSTM model has a stronger advantage for the sequence annotation task due to its memory function. The CRF model makes full use of the relationship of adjacent tags based on the BILSTM model to optimize the optimal splicing of the whole sequence. ALBERT-BILSTM-CRF model has a higher F_1 value compared with the BILSTM-CRF model and

ALBERT-CRF model F_1 values are 3.63% and 0.59% higher, which can better express the semantic information of words because the word vectors generated by the ALBERT pre-trained language model are contextually relevant. In this paper, the multi-task learning model is also trained on the ALBERT embedding layer and BILSTM encoding layer during the intent classification training and our method can effectively improve the named entity recognition effect compared with the ALBERT-BILSTM-CRF model without multi-task learning strategy.

The enhancement of named entity recognition effect can be explained by the intent classification task providing additional training data and the multi-task learning is equivalent to an implicit data enhancement. By sharing the underlying parameters, the rich semantic and association information in the intent recognition task is learned, and the generalization ability of the model is improved.

From the recognition effect of each entity type, the multi-task learning model in this paper also stands out in the F_1 value as shown in Figure 4. The F_1 values of our model are higher than other models in all three entity types, which validates the effectiveness of our method.

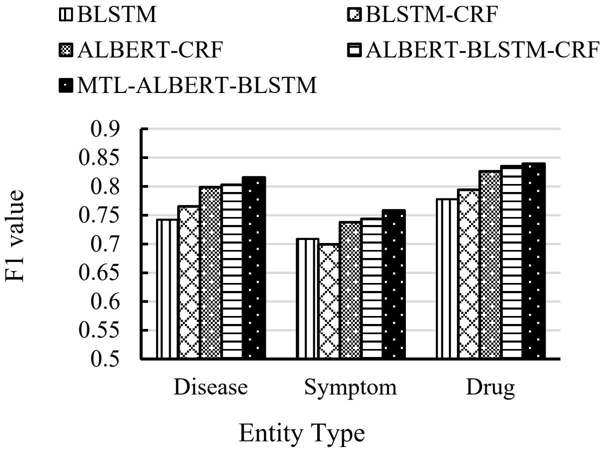

Figure 4. Comparison of F_1 values of each model in the three entity types.

For the intent classification task, we used F_1 values, precision and recall to evaluate our model. We used CNN [13], BILSTM and ALBERT [21] as benchmark methods, and we also compared with the single-task model for intent recognition, ALBERT-BILSTM, and the experimental results are shown in Table 3.

Table 3. Results of intent classification on different model.

Method	P	R	F_1
CNN	0.8315	0.7921	0.8113
BILSTM	0.8377	0.8086	0.8229
ALBERT	0.8582	0.8391	0.8485
ALBERT-BILSTM	0.8634	0.8473	0.8553
MTL-ALBERT-BILSTM	**0.8842**	**0.8654**	**0.8747**

Our model also outperforms other benchmark model approaches in terms of F_1 value, precision and recall in intent classification task. In addition, our multi-task learning model shows a significant improvement over the ALBERT-BILSTM single-task model, which means that the multi-task learning approach significantly enhances the intent classification capability. The performance improvement of intent classification is more pronounced than named entity recognition, and the F_1 value of the intent classification task is about 2% higher than that of the ALBERT-BILSTM model using a single-task learning strategy.

Intent classification is a less complex task in that it only needs to generate labels for the entire sentence unlike the named entity recognition task which generates labels for each word. The model can learn more semantic information in medical named entity labeling, using the entity recognition ability trained by the named entity recognition task, which explains to some extent the significant improvement in the intention classification task. In multi-task learning, each task can "selectively" use the hidden features learned in other tasks to improve its capabilities.

In addition, it can be seen from Figure 5 that the classification performance of the multi-task model in this paper is better than other methods in all seven intent categories, which verifies the feasibility of the ALBERT-BILSTM-based multi-task learning model in this paper.

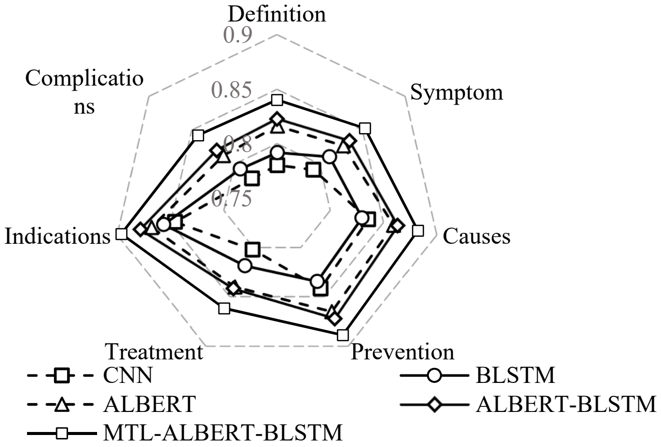

Figure 5. Comparison of the F_1 value of different methods in each intent classifications.

5. Conclusions and Future Work

In this paper, an ALBERT-BILSTM-based multi-task learning model is proposed for Chinese medical questions intent classification and named entity recognition tasks. Comparative experiments of different models are conducted on the Chinese medical questions dataset. The experimental results show that pre-trained language models and multi-task learning strategies can work together to improve natural language understanding of medical texts. Compared with the benchmark approach and the single-task model, the generalization ability of the model is well improved. The named entity recognition and intent classification task share the ALBERT word embedding layer and BILSTM layer. The model outputs named entity labels or intent labels using their task-specific layers. During multi-task training, parameters in the word embedding layer and BILSTM layer can be updated simultaneously by these tasks, facilitating each other's learning to obtain rich semantic and associative information at word level and sentence level. There are also some recent works that investigate learning which parameters can be shared, and these works outperform the hard sharing mechanism in a general sense. In future work we try to investigate task-specific goals and inter-task optimization tradeoffs to make the shared representations learned by the model more accurate and thus further improve the predictive performance of the model.

Author Contributions: Conceptualization, T.T. and M.A.; methodology, M.A. and T.T.; software, M.A.; validation, T.T., M.A. and A.H.; formal analysis, T.T.; investigation, M.A.; resources, M.A.; data curation, T.T.; writing—original draft preparation, M.A.; writing—review and editing, T.T.; visualization, A.H.; supervision, T.T.; project administration, T.T.; funding acquisition, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62166042, U2003207), Natural Science Foundation of Xinjiang, China (2021D01C076), and Strengthening Plan of National Defense Science and Technology Foundation of China (2021-JCJQ-JJ-0059).

Data Availability Statement: The dataset supporting the conclusions of this article is available at https://github.com/zhangsheng93/cMedQA2 and http://www.github.com/liutongyang/CMID, accessed on 28 October 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gerner, M.; Nenadic, G.; Bergman, C.M. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinform.* **2010**, *11*, 85. [CrossRef] [PubMed]
2. Fukuda, K.I.; Tsunoda, T.; Tamura, A.; Takagi, T. Toward information extraction: Identifying protein names from biological papers. *Pac. Symp. Biocomput.* **1998**, *707*, 707–718.
3. He, L.; Yang, Z.; Lin, H.; Li, Y. Drug name recognition in biomedical texts: A machine-learning-based method. *Drug Discov. Today* **2014**, *19*, 610–617. [CrossRef] [PubMed]
4. Chen, Y.; Zhou, C.; Li, T.; Wu, H.; Zhao, X.; Ye, K.; Liao, J. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. *J. Biomed. Inform.* **2019**, *96*, 103252. [CrossRef] [PubMed]
5. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tag-grog. *arXiv* **2015**, arXiv:1508.01991.
6. Yang, P.; Yang, Z.; Luo, L.; Lin, H.; Wang, J. An attention-based approach for chemical compound and drug named entity recognition. *J. Comput. Res. Dev.* **2018**, *55*, 1548–1556.
7. Li, L.; Guo, Y. Biomedical named entity recognition with CNN-BILSTM-CRF. *J. Chin. Inf. Process.* **2018**, *32*, 116–122.
8. Su, Y.; Liu, J.; Huang, Y. Entity Recognition Research in Online Medical Texts. *Acta Sci. Nat. Univ. Pekin.* **2016**, *52*, 1–9.
9. Qin, Q.; Zhao, S.; Liu, C. A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records. *Complexity* **2021**, *2021*, 6631837. [CrossRef]
10. Ji, B.; Li, S.; Yu, J.; Ma, J.; Tang, J.; Wu, Q.; Tan, Y.; Liu, H.; Ji, Y. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models. *J. Biomed. Inform.* **2020**, *104*, 103395. [CrossRef]
11. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40. [CrossRef]
12. Ravuri, S.; Stolcke, A. A comparative study of recurrent neural network models for lexical domain classification. In Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 6075–6079.
13. Zhang, S.; Grave, E.; Sklar, E.; Elhadad, N. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *J. Biomed. Inform.* **2017**, *69*, 1–9. [CrossRef]
14. Yao, L.; Mao, C.; Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 31–39. [CrossRef]
15. Jang, B.; Kim, M.; Harerimana, G.; Kang, S.U.; Kim, J.W. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Appl. Sci.* **2020**, *10*, 5841. [CrossRef]
16. Zhang, Q.; Yuan, Q.; Lv, P.; Zhang, M.; Lv, L. Research on Medical Text Classification Based on Improved Capsule Network. *Electronics* **2022**, *11*, 2229. [CrossRef]
17. Zaib, M.; Sheng, Q.Z.; Emma Zhang, W. A short survey of pre-trained language models for conversational ai-a new age in nlp. In Proceedings of the Australasian Computer Science Week Multiconference, Melbourne, VIC, Australia, 4–6 February 2020; pp. 1–4.
18. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
19. Song, Z.; Xie, Y.; Huang, W.; Wang, H. Classification of traditional chinese medicine cases based on character-level bert and deep learning. In Proceedings of the 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 24–26 May 2019; pp. 1383–1387.
20. Yao, L.; Jin, Z.; Mao, C.; Zhang, Y.; Luo, Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1632–1636. [CrossRef]
21. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
22. Zhang, Z.; Jin, L. Clinical short text classification method based on ALBERT and GAT. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022; pp. 401–404.
23. Yang, Q.; Shang, L. Multi-task learning with bidirectional language models for text classification. In Proceedings of the International Joint Conference on Neural Network (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
24. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
25. Liu, P.; Qiu, X.; Huang, X. Recurrent neural network for text classification with multi-task learning. *arXiv* **2016**, arXiv:1605.05101.
26. Wu, Q.; Peng, D. MTL-BERT: A Multi-task Learning Model Utilizing Bert for Chinese Text. *J. Chin. Comput. Syst.* **2021**, *42*, 291–296.

27. Chowdhury, S.; Dong, X.; Qian, L.; Li, X.; Guan, Y.; Yang, J.; Yu, Q. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. *BMC Bioinform.* **2018**, *19*, 75–84. [CrossRef] [PubMed]
28. Zhao, S.; Liu, T.; Zhao, S.; Wang, F. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In Proceedings of the AAAI Conference on Artificial Intelligence, Budapest, Hungary, 27 January–1 February 2019; Volume 33, pp. 817–824. [CrossRef]
29. Peng, Y.; Chen, Q.; Lu, Z. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. *BioNLP* **2020**, *2020*, 205.
30. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 5753–5763.
31. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
32. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Hochreiter, S.; Schmidhuber, J. Long short memory. *Neural Comput.* **2014**, *9*, 1735–1780. [CrossRef]
35. Zhang, S.; Zhang, X.; Wang, H.; Cheng, J.; Li, P.; Ding, Z. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. *Appl. Sci.* **2017**, *7*, 767. [CrossRef]
36. Chen, N.; Su, X.; Liu, T.; Hao, Q.; Wei, M. A benchmark dataset and case study for Chinese medical question intent classification. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 125. [CrossRef]