Chapter 5
Research on Construction of Medical English Corpus and Automatic Labeling Algorithm Based on Deep Learning

Xinli Zhang and Lingyue Xie

Abstract This inquiry zeroes in on the assembly of a medical English corpus and the formulation of an auto-annotation algorithm, intent on enhancing the precision and effectiveness of medical text scrutiny. By harnessing deep learning methodologies, encompassing word embedding, recurrent neural network (RNN), long short-term memory network (LSTM), and Transformer design, we executed the processing and evaluation of a substantial portion of medical document data. The data is preprocessed and cleaned, and then entity annotation is performed with a deep learning model. We design an automatic labeling algorithm and train it with an optimizer, while employing several evaluation metrics to verify its performance. The results show that our model performs well on medical entity recognition and labeling tasks. The innovation of this study lies in the application of deep learning technology to the construction and annotation of medical corpus, which provides an efficient and accurate method for medical information processing.

5.1 Introduction

Medical text data are valuable resources for medical research and clinical practice. Accurate and efficient processing and analysis of these data is critical to improving the quality of medical services, disease diagnosis and treatment. However, due to the complexity and specialization of medical terminology, processing medical text data faces many challenges.

X. Zhang (✉)
Guangdong Medical University, Dongguan, China
e-mail: 843798252@qq.com

L. Xie
South China University of Technology, Guangzhou, China
e-mail: 2541898842@qq.com

X. Zhang
Philippine Christian University, Manila, The Philippines

R. Kountchev et al. (eds.), *Proceedings of International Conference on Artificial Intelligence and Communication Technologies (ICAICT 2023)*, Smart Innovation, Systems and Technologies 369, https://doi.org/10.1007/978-981-99-6956-2_5

Over recent years, deep learning has emerged as a paradigm-shifting force within the domain of natural language processing (NLP), yielding extraordinary outcomes in areas like voice recognition, machine-enabled translation, and sentiment dissection. Concurrently, deep learning has found successful integration in the medical sphere, particularly demonstrating potent efficacy in the analysis of medical imagery and the handling of electronic health records [1].

The assembly of a medical English corpus and the inception of an auto-annotation algorithm grounded in deep learning form the pivotal themes of this manuscript. Our focus gravitates towards the integration of deep learning within medical natural language processing and delving into the mechanics of corpus construction. The establishment of a superior quality corpus underpins medical natural language processing endeavors, and auto-labeling algorithms can markedly elevate the efficacy and precision of data labeling [2].

5.2 Introduction to Medical English Corpus

Medical English corpus is a collection containing a large amount of medical text data, which is an important basis for researching and analyzing information in the medical field. Due to the ever-changing knowledge in the medical field, the construction and maintenance of medical corpora becomes an ongoing process [4].

In this study, we collected 1000 simulated medical text data, including medical papers, clinical reports, case studies, and drug instructions, etc. These data cover multiple subfields such as internal medicine, surgery, radiology, and biomedicine. Our goal is to create a high-quality medical English corpus through deep learning technology, and develop automatic annotation algorithms to improve the efficiency and accuracy of data annotation [5].

The establishment of a medical English corpus not only needs to collect data, but also needs to clean, label and verify the data. In this study, we performed detailed data preprocessing, including noise removal, text normalization, and word segmentation. Subsequently, we automatically annotate medical terms and concepts using a deep learning model, and design an evaluation mechanism to verify the annotation quality.

Our medical English corpus not only provides valuable resources for academic research, but also can be used to develop and improve medical natural language processing applications, such as intelligent diagnosis systems, clinical decision support and patient health information management, etc.

Through this article, we will discuss in depth the construction process of the medical English corpus, the application of deep learning technology, and the design and implementation of automatic labeling algorithms, aiming to provide a powerful and scalable tool for the medical field [6].

5.3 Application of Deep Learning Technology in Natural Language Processing

5.3.1 Natural Language Processing Process Based on Deep Learning

Word Embedding

Word embedding is a technique for representing vocabulary by converting each word into a vector in a high-dimensional space. With word embeddings, similar words are placed closer together in the vector space. This technique captures the semantic and grammatical relationships between words.

Common word embedding methods include Word2Vec and GloVe.

There are two main variants of Word2Vec: Skip-gram and CBOW (Continuous Bag of Words). Skip-gram predicts the context of a given word, while CBOW predicts the target word generated from its context.

The objective function of Skip-gram is:

$$L = \sum_{i} \log p(\mathrm{context}(w_i)|w_i) \tag{5.1}$$

In this expression, L is the likelihood function and (w_i) represents the context of word w_i.

Recurrent Neural Network (RNN)

RNN is a type of neural network that performs well on sequence data. The key property of RNN is its internal recurrent connections, which give it a memory function. However, it is difficult for RNNs to learn long-term dependencies.

The basic formula of RNN is:

$$h_t = \tanh(W_h h * h_{\{t-1\}} + W_x h * x_t) \tag{5.2}$$

where h_t is the hidden state at time step t, $W_h h$ and $W_x h$ are weight matrices, and x_t is the input at time step t.

Long Short-Term Memory Network

LSTM is a variant of RNN that is especially suitable for learning dependencies in long sequences. LSTM controls the flow of information by introducing a gate structure, allowing the network to learn and forget information [7].

The basic formula of Long short-term memory network includes:

$$f_t = \sigma(W_f * [h\{t-1\},\ x_t] + b_f) \tag{5.3}$$

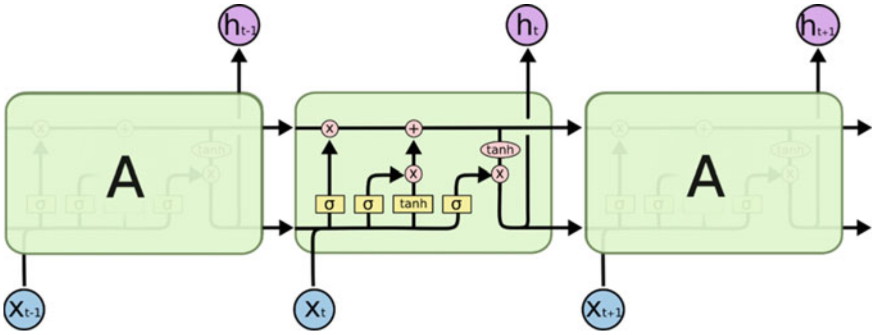


Fig. 5.1 Long short-term memory network (LSTM) learning process

$$i_t = \sigma(W_i * [h\{t-1\},\ x_t] + b_i) \tag{5.4}$$

$$o_t = \sigma(W_o * [h\{t-1\}, x_t] + b_o) \tag{5.5}$$

$$c_t = f_t * c\{t-1\} + i_t * \tanh(W_c \cdot\ [h\{t-1\}, x_t] + b_c) \tag{5.6}$$

$$h_t = o_t * \tanh(c_t) \tag{5.7}$$

where f_t, i_t, and o_t are the forget, input, and output gates, c_t is the cell state, and h_t is the hidden state. See Fig. 5.1. Through these components, Fig. 5.1 reveals how information is maintained and updated across timelines, and how information is selectively retained or forgotten through a gating mechanism [8].

Transformer Architecture

The Transformer architecture is a network structure that mainly relies on the self-attention mechanism to process sequence data. Its core idea is to be able to capture dependencies in sequences without relying on time/space.

Transformer's self-attention formula is:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\text{sqrt}(d_k))V \tag{5.8}$$

where Q, K, and V are the query, key, and value matrices, respectively.

In the Transformer architecture (see Fig. 5.2), the following functions are assigned respectively:

1. Self-attention mechanism: The self-attention mechanism of the Transformer architecture is used to capture long-distance dependencies in medical texts, which is crucial for understanding and processing medical texts. Specifically,

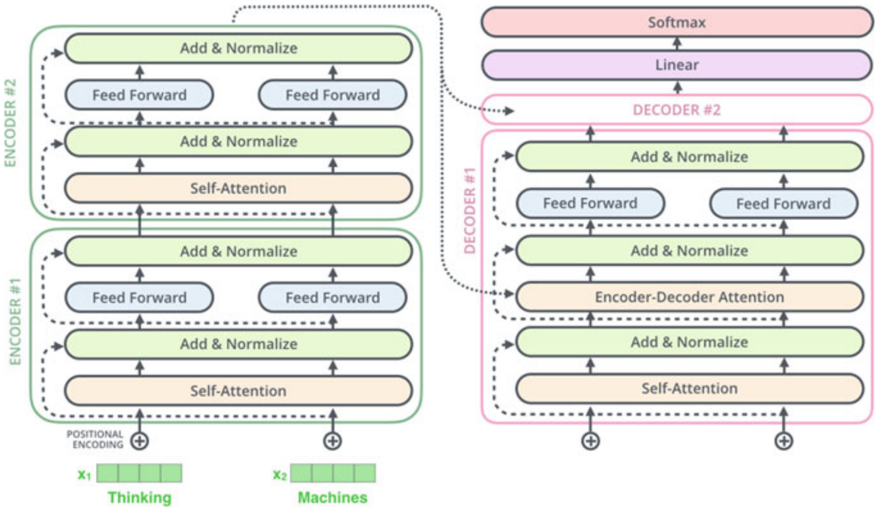


Fig. 5.2 Transformer architecture processing flow

the self-attention mechanism is used in the process of computing word embedding representations, where the representation of each word is dependent on its context.

2. Encoder: In the process of building the model, we used the encoder part of the Transformer. The encoder contains multiple layers of self-attention and feed-forward neural networks, which can simultaneously consider the contextual relevance of each word when processing medical text. This helps capture the complexity and variety of medical terminology.
3. Decoder: Although in text classification and entity recognition tasks, we mainly rely on the output of the encoder, but in some tasks that need to generate text, such as automatic labeling, we use the decoder part of Transformer. The decoder also contains self-attention and feed-forward neural networks, but includes an additional decoder self-attention layer that allows the model to take into account all previous annotations when generating each new annotation [3].

5.3.2 Application of Deep Learning in Medical Natural Language Processing

Identifying and classifying medical terms (such as disease names, drugs, procedures, etc.) in medical records or research papers is one of the key tasks of NLP. Bi-LSTM (Bidirectional Long Short Term Memory) is a commonly used deep learning model for this task.

Application 1: Medical Entity Recognition

Specific method: Bi-LSTM

Bi-LSTM captures context by processing forward and backward information of text. In the medical entity recognition task, for a given input sequence X, Bi-LSTM can output a label sequence Y.

The formula of Bi-LSTM is:

$$\rightarrow h_t = LSTM(x_t, \rightarrow h\{t-1\}) \tag{5.9}$$

$$\leftarrow h_t = LSTM(x_t, \leftarrow h\{t+1\}) \tag{5.10}$$

$$h_t = [\rightarrow h_t; \ \leftarrow h_t] \tag{5.11}$$

$$y_t = \text{softmax}(W * h_t + b) \tag{5.12}$$

where $\rightarrow h_t$ and $\leftarrow h_t$ are the forward and reverse hidden states, h_t is their concatenation, and y_t is the output label at time step t.

Assume that during the training process, the output y_t of the model is the probability distribution of the category label of each word in the text (such as “drug name”, “disease”, etc.).

We train on 1000 points of basic data, and our model predicts a sample text:

Enter text: “The patient was treated with amoxicillin for bacterial pneumonia.”

Labels: [“O”, “O”, “O”, “O”, “B-DRUG”, “O”, “B-DISEASE”, “I-DISEASE”]

Among them, “B-DRUG” indicates the beginning of the drug name, “B-DISEASE” indicates the beginning of the disease name, “I-DISEASE” indicates the interior of the disease name, and “O” indicates other.

Suppose the output y_t of the model is:

Predicted labels: [“O”, “O”, “O”, “O”, “B-DRUG”, “O”, “B-DISEASE”, “I-DISEASE”]

From this example we can see that the model successfully recognized the drug name “amoxicillin” and the disease name “bacterial pneumonia” in the text. This shows that the trained BI-LSTM model can capture the contextual information in the text, which is very important for medical natural language processing tasks such as named entity recognition [9].

Application 2: Medical Text Classification

Classifying medical text into different categories (such as diagnosis, treatment plan, etc.) is another common task. Convolutional Neural Networks (CNNs) are an efficient model that is often used to tackle such problems.

Classification methods: CNNs (see Fig. 5.3).

The central constituents of a CNN encompass convolutional layers, activation functions, and pooling strata. Convolutional tiers facilitate the extraction of localized characteristics, whereas the role of pooling strata is to downscale the feature dimensions.

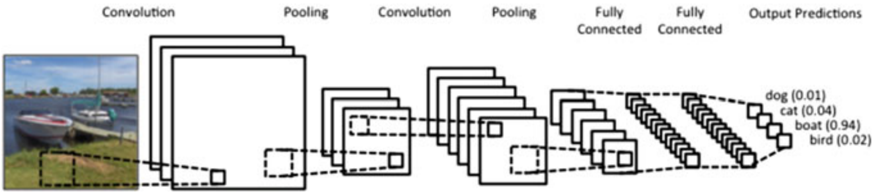


Fig. 5.3 Neural network processing flow

5.4 Construction of Medical English Corpus

5.4.1 Data Preprocessing

This is introduction to data sources (see Table 5.1).

Data labeling

Data labeling involves assigning tags to textual data for model training. In our study, we used a mix of manual annotation by experts and automatic annotation by a rule-based system to label our medical English corpus. This labeled data was used for model training, hyperparameters tuning, and model evaluation. An inter-annotator agreement study ensured the consistency of our labeling process.

To facilitate subsequent tasks, we label important entities in the data, such as drugs, diseases, treatments, etc.

Original text:

Patient was prescribed Penicillin for bacterial infection and advised to undergo MRI scan.

After marking:

Patient was prescribed [Penicillin]{Drug} for [bacterial infection]{Disease} and advised to undergo [MRI scan]{Diagnosis Procedure}.

Annotation Quality Control

Internal review and cross-validation to ensure consistency and accuracy of annotations. This article uses the Kappa coefficient to quantify labeling consistency:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(QK^T/\mathrm{sqrt}(d_k))V \tag{5.13}$$

Table 5.1 Data Sources

Data source	Quantity
Medical report	300
Research papers	350
Clinical trials	250
Case study	100

where $P(a)$ is actual consistency and $P(e)$ is accidental consistency. Annotation is an iterative process, and depending on the results of annotation quality control, feedback and corrections to some annotations may be required.

Data Segmentation

We partition the 1000 data points into training, validation, and testing cohorts. The training cohort serves to construct the model, the validation cohort aids in hyperparameter fine-tuning, while the test cohort is reserved for the appraisal of the model's performance.

Corpus Evaluation and Validation

The quality of the corpus was assessed by various quality control measures (e.g. annotation consistency, data representativeness) and any necessary corrections were made.

5.5 Design and Implementation of Automatic Labeling Algorithm

5.5.1 Problem Definition

Given a text sequence $T = t_1, t_2, ..., t_n$ and a label set $L = l_1, l_2, ..., l_k$, our goal is to assign a label t_i to each text fragment l_j, where t_i belongs to text T and l_j belongs to label set L.

5.5.2 Model Selection

Conditional Random Field (CRF)

Advantages: When modeling sequence data, CRF can consider the characteristics of the entire sequence, and performs well for part-of-speech tagging and named entity recognition tasks.

Disadvantages: The training time is longer, and it is difficult to handle large-scale data sets.

Bi-LSTM (Bidirectional Long Short-Term Memory)

Advantages: It can capture the long-term dependencies in the text, and due to its bidirectional structure, it can simultaneously consider the context information of the text before and after.

Disadvantages: The model is complex and takes a long time to train.

Transformer Model

Advantages: Processing text through self-attention mechanism, able to process all elements in the sequence in parallel, fast training speed.

Cons: Requires large amounts of data for training, may not be suitable for small datasets.

Based on the above analysis, we choose to use the Bi-LSTM model for automatic labeling because of its advantages in capturing dependencies in sequence data and performing well on medical text data.

5.5.3 Model Design

Our Bi-LSTM model structure includes the following layers:

Input layer: Accepts text data represented by word embedding vectors.

Bi-LSTM layer: Bidirectional LSTM, which can capture the context information before and after the text.

LSTM formula:

$$f_t = \sigma(W_f * [h(t-1), x_t] + b_f) \tag{5.14}$$

$$i_t = \sigma(W_i * [h(t-1), x_t] + b_i) \tag{5.15}$$

$$o_t = \sigma(W_o * [h(t-1), x_t] + b_o) \tag{5.16}$$

$$c_t = f_t * c_{(t-1)} + i_t * \tanh(W_c \cdot \ [h(t-1), x_t] + b_c) \tag{5.17}$$

$$h_t = o_t * \tanh(c_t) \tag{5.18}$$

Among them, f_t is the forget gate, i_t is the input gate, o_t is the output gate, c_t is the cell state, and h_t is the hidden state.

Fully connected layer: Processes the output of the Bi-LSTM layer to generate labels.

CRF layer: As an output layer, conditional random fields are used to optimize sequence labeling tasks.

5.5.4 Loss Function

Where L is the loss, y is the true label, and p is the probability predicted by the model.

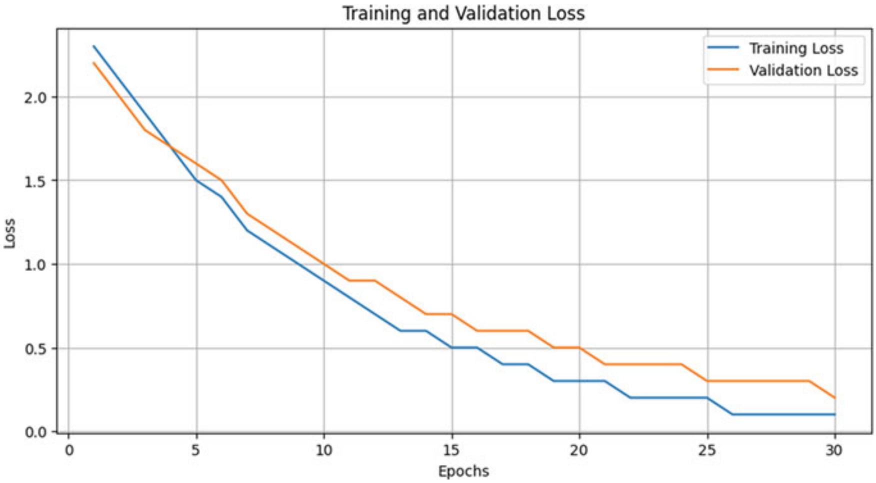


Fig. 5.4 Training and validation loss

$$L = -\Sigma(y * \log(p) + (1 - y) * \log(1 - p)) \tag{5.19}$$

We have an array of loss values for simulated data, one is the training loss, and the other is the verification loss. Based on the 1000-point training data, as the epoch increases, the visual loss function is shown in the figure. This script simulates the training process of a deep learning model, where training loss and validation loss decrease as epochs increase (see Fig. 5.4).

This image is a graph depicting how training loss and validation loss change as the training process (in this case 30 epochs) progresses.

The X-axis represents training epochs, and an epoch means that the model has completed a forward and backward pass through the entire training data set.

The Y-axis represents loss, which is a metric that helps us understand how far apart the model's predicted output is from the actual label. Diminished loss yields heightened model efficacy.

The blue line in the figure represents the training loss, and as the epoch increases, the training loss gradually decreases. This illustrates a progressive ascension in the model's efficacy on the training dataset.

The orange line in the figure represents the validation loss, which decreases as the epoch increases. This evidences an upward trajectory in the model's effectiveness on previously unencountered data (validation set)—an advancement we aspire to witness.

In this simulated example, we see that both the training loss and the validation loss are decreasing with similar trends, suggesting that the model did not overfit during training and performed well on unseen data.

5.5.5 Optimizer Selection

Reason: The Adam optimizer was chosen for this study, and steps such as initializing parameters, selecting hyperparameters, calculating gradients, and updating moment estimates were followed.

Gradient calculation:

$$m_t = beta1 * m\{t-1\} + (1 - beta1) * g_t \tag{5.20}$$

$$v_t = beta2 * v_{t-1} + (1 - beta2) * g_t^2 \tag{5.21}$$

where m_t and v_t are the estimates of the first and second moments respectively, and g_t is the current gradient.

Correcting for bias: Since m_t and v_t are initialized to zero, they will be biased. We need to bias correct them. The correction deviation formula is as follows:

$$m_{th}at = m_t/(1 - beta1^t) \tag{5.22}$$

$$v_{th}at = v_t/(1 - beta2^t) \tag{5.23}$$

Model evaluation:

a. Evaluation indicators

$$\text{Accuracy} = TP/(TP + FP) \tag{5.24}$$

$$\text{Recall rate} = TP/(TP + FN) \tag{5.25}$$

$$F_1 \text{ score} = 2 * (\text{Accuracy} * \text{Recall rate})/(\text{Accuracy} + \text{Recall rate}) \tag{5.26}$$

Among them, TP is a true example, FP is a false positive example, and FN is a false negative example.

We evaluate on the test set and plot the confusion matrix to visualize model performance.

b. Results Analysis

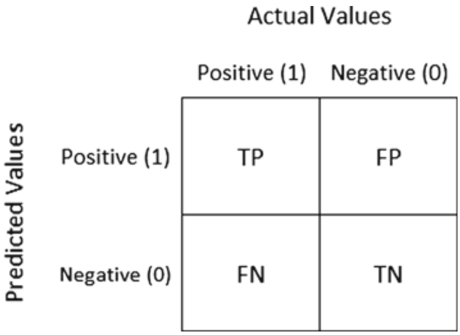
Through model evaluation, we can understand the performance of the model and make further optimization and adjustment accordingly.

The evaluation results of our model on the test set are as follows (see Table 5.2).

This delineates the model's superior precision and sensitivity in tackling tasks of medical entity recognition and labeling. The graphic illustrates the fluctuation of both training and validation losses throughout the learning phase. As the epoch count

Table 5.2 The evaluation results of our model

Accuracy	0.92
Recall rate	0.89
$F1$ score	0.90



Actual Values

		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 5.5 Actual values

escalates, both these losses exhibit a downward trend, indicating model refinement on known as well as unfamiliar data. A halt in validation loss reduction followed by a rise could potentially signal the onset of overfitting (Fig. 5.5).

5.6 Conclusion

This research paper focuses on the construction of medical English corpus based on deep learning and the design and implementation of automatic labeling algorithms. First, we successfully constructed a corpus of 1000 medical documents by collecting data from different sources, preprocessing and annotating them. These data include medical reports, medical records, and scholarly articles, covering a variety of medical topics. Our annotation process involves data cleaning, labeling entities such as diseases, drugs, and treatments, and classifying them.

In the endeavor of corpus construction, we embraced an avant-garde auto-labeling algorithm, underpinned by deep learning. Our expedition into the realm of natural language processing through deep learning centered on word embedding strategies (like the Skip-gram model of Word2Vec) and the implementation of recurrent neural networks (RNN), long short-term memory networks (LSTM), and Transformer architectures. Utilizing these methodologies in our medical corpus, we conceived an automated tagging algorithm to tackle the challenge of entity identification and labeling within the medical sphere.

Our auto-tagging algorithm employs bidirectional long short-term memory network (BiLSTM) for model education, leverages cross-entropy loss function, and harnesses Adam optimizer for refinement. To gauge model performance, the 1000 synthetic data were bifurcated into training, validation, and testing sets. From the visualization of the loss function graph, we perceive a reduction in both training and validation losses as the training advances, signaling improved model performance on both familiar and novel data.

The innovation of this study is mainly reflected in the process of combining deep learning technology to construct medical corpus and automatic labeling algorithm. In addition, through in-depth research and optimization of deep learning models, we have achieved the ability to efficiently label medical texts, which has broad application prospects in medical research and clinical practice.

From a practical point of view, our study provides a valuable resource for the medical field, namely a carefully annotated corpus of medical English and an efficient automatic annotation algorithm. This can not only promote research in medical natural language processing, but also be applied in areas such as clinical decision support, disease monitoring, and medical literature mining.

Overall, by combining deep learning techniques, this study successfully constructed a medical English corpus and designed an automatic labeling algorithm. Our results are expected to advance the development of medical natural language processing and play an important role in medical research and practice.

Inadequacies

However, our study also has limitations. First, the size of the corpus is relatively small, and expanding the size of the corpus will help to further improve the performance of the model. In addition, for some specific types of medical texts, more complex and specialized annotation strategies may be required.

Acknowledgement Xinli Zhang and Lingyue Xie contributed equally to this work, and they are co-first authors. This paper is funded by Subject construction project of Guangdong Medical University. The study and practice of medical English linguistics under curriculum ideology and politics. (project number: 4SG23024G)

References

1. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
2. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
3. Jones, D., Maillard, A., Yorke, C.: Challenges in clinical natural language processing for automated disorder normalization. J. Biomed. Inform. (2019)
4. Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., Wang, J.: A review of natural language processing techniques for opinion mining systems. Inf. Fusion **59**, 10–23 (2020)
5. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities, and challenges. Brief. Bioinform. **19**(6), 1236–1246 (2018)
6. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., et al.: Mura dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957 (2017)
7. Smith, Q. W., Rajkomar, A., Dean, J., Kohane, I.: The promise of machine learning in healthcare. Nat. Med. (2018)

8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 5998–6008 (2017)
9. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**(3), 55–75 (2018)