International Journal of Medical Informatics 132 (2019) 103985

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Extracting comprehensive clinical information for breast cancer using deep learning methods

Xiaohui Zhang[a,1], Yaoyun Zhang[b,1], Qin Zhang[b], Yuankai Ren[b], Tinglin Qiu[c], Jianhui Ma[c,*], Qiang Sun[a,*]

[a] Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical Sciences, Beijing, China
[b] Digital China Health Technologies Co. Ltd., Beijing, China
[c] National Cancer Center/Cancer Hospital, Peking Union Medical College & Chinese Academy of Medical Sciences, Beijing, China

ARTICLE INFO

Keywords:
Clinical information extraction
Breast cancer
Deep learning
Fine-tuning BERT
Information model

ABSTRACT

Objective: Breast cancer is the most common malignant tumor among women. The diagnosis and treatment information of breast cancer patients is abundant in multiple types of clinical fields, including clinicopathological data, genotype and phenotype information, treatment information, and prognosis information. However, current studies are mainly focused on extracting information from one specific type of clinical field. This study defines a comprehensive information model to represent the whole-course clinical information of patients. Furthermore, deep learning approaches are used to extract the concepts and their attributes from clinical breast cancer documents by fine-tuning pretrained Bidirectional Encoder Representations from Transformers (BERT) language models.

Materials and methods: The clinical corpus that was used in this study was from one 3A cancer hospital in China, consisting of the encounter notes, operation records, pathology notes, radiology notes, progress notes and discharge summaries of 100 breast cancer patients. Our system consists of two components: a named entity recognition (NER) component and a relation recognition component. For each component, we implemented deep learning-based approaches by fine-tuning BERT, which outperformed other state-of-the-art methods on multiple natural language processing (NLP) tasks. A clinical language model is first pretrained using BERT on a large-scale unlabeled corpus of Chinese clinical text. For NER, the context embeddings that were pretrained using BERT were used as the input features of the Bi-LSTM-CRF (Bidirectional long-short-memory-conditional random fields) model and were fine-tuned using the annotated breast cancer notes. Furthermore, we proposed an approach to fine-tune BERT for relation extraction. It was considered to be a classification problem in which the two entities that were mentioned in the input sentence were replaced with their semantic types.

Results: Our best-performing system achieved F1 scores of 93.53% for the NER and 96.73% for the relation extraction. Additional evaluations showed that the deep learning-based approaches that fine-tuned BERT did outperform the traditional Bi-LSTM-CRF and CRF machine learning algorithms in NER and the attention-Bi-LSTM and SVM (support vector machines) algorithms in relation recognition.

Conclusion: In this study, we developed a deep learning approach that fine-tuned BERT to extract the breast cancer concepts and their attributes. It demonstrated its superior performance compared to traditional machine learning algorithms, thus supporting its uses in broader NER and relation extraction tasks in the medical domain.

1. Introduction

Breast cancer is the most common malignant tumor among women, with more than 2 million new cases and more than 600,000 deaths occurring every year worldwide. In the United States, more than 200,000 new cases occur every year with more than 40,000 deaths [1,2]. According to the latest annual report on cancer in China, 279,000 new breast cancer cases and 69,000 breast cancer related deaths occurred in 2014. The incidence of breast cancer in China has rapidly increased in recent years. The incidence of breast cancer has reached 41.82 per 100,000 person-years [3].

Breast cancer patients have abundant diagnosis and treatment

* Corresponding authors.
E-mail addresses: majianhui@csco.org.cn (J. Ma), sunqpumch@163.com (Q. Sun).
[1] First author that contributed equally.

https://doi.org/10.1016/j.ijmedinf.2019.103985
Received 11 September 2019; Accepted 29 September 2019

information, including clinicopathological data, genotype and phenotype information, treatment information and prognosis follow-up information. The systematic summary and resource sharing of this tremendous diagnosis and treatment information are important means to better manage patients, explore new diagnosis and treatment methods, and implement accurate medical treatments. Fortunately, the widely disseminated electronic health record systems (EHRs) have accumulated a large amount of longitudinal clinical information. EHRs have also contributed to many other secondary clinical applications, such as measuring medical quality and predicting hospital readmissions.

One essential challenge to the secondary uses of EHRs is that a large portion of useful clinical information is recorded as unstructured clinical text. Therefore, natural language processing (NLP) methods have been applied to extract different types of information from breast cancer notes. Several previous studies have validated the feasibility of using machine learning-based NLP to extract clinical information from breast pathology reports, such as the procedures and results [4–6]. Rule-based and/or machine learning-based NLP methods were also applied to mammography reports to extract information such as findings and BI-RADS assessment categories [7–9]. More recently, researchers began to use deep learning methods, e.g., convolutional neural networks (CNN), to extract the primary sites from breast cancer pathology reports [10].

Although promising results have been obtained by multiple information extraction systems for breast cancer in the US, only a few efforts have focused on Chinese clinical text. One potential reason is that clinical NLP research is still in its infancy in China. In our previous work, a deep learning-based Bi-LSTM-CRF algorithm was used to extract information on BI-RADS assessment categories from Chinese mammography reports, which outperformed the classic CRF algorithm and a lexicon-based baseline method [11].

Moreover, current works are mainly focused on extracting information from pathology and radiology reports. As mentioned above, the clinical information of breast cancer patients is abundant, including a wide range of information including demographic information such as the professions of patients, medical history information such as the age of a mother's first child, family history information such as the cancer history of one's parents, clinicopathological information such as pathological types, and biomarker information such as Her-2. Unfortunately, there is no clinical IE system that can interpret a comprehensive set of clinical breast cancer characteristics in the biomedical community so far.

This study takes the initiative to design a breast cancer information model representing a comprehensive collection of clinical concepts and their relations. Moreover, we investigated the feasibility of using BERT [12], the latest deep learning language model based on the attention mechanism, for the information extraction from Chinese clinical breast cancer notes. BERT has achieved the optimal performance on NLP tasks for both open text and English clinical text, in comparison with current state-of-the-art machine learning methods [12,13]. It performs better than previous methods because it applies the first attention-based bidirectional neural network architecture to jointly represent the sequence information in both directions for language models. A clinical language model is first trained using BERT on a large unlabeled corpus of Chinese clinical text. Next, a clinical IE system is built by fine-tuning the pretrained language model to extract both the named entities and their relations from the encounter notes, operation records, pathology notes, radiology notes, progress notes and discharge summaries of hospitalized breast cancer patients. The experimental results demonstrated that the proposed methods outperformed several state-of-the-art deep learning and classic machine learning algorithms on both named entity recognition (NER) and relation extraction tasks.

2. Method

2.1. Dataset

As mentioned above, the dataset contains six types of clinical records, including the encounter notes, operation records, pathology notes, radiology notes, progress notes and discharge summaries of 100 breast cancer inpatients. In total, the dataset consists of 8473 sentences. The selection criteria of the notes are described as follows.

Inclusion criteria: The patient should have clinical records of their primary breast cancer surgery. The patient's records should have all six types of clinical notes that are processed in this study, which will contain clinical information on the potential cause, diagnosis, treatment, and prognosis of their breast cancer. Exclusion criteria: Patients with historical records of other types of cancer are excluded. Patients who do not have a complete set of note types that are processed in this study are also excluded. Since this study mainly focused on building a high-performance breast cancer information extraction system, there was no further constraints such as age limits for selecting the dataset, as in a cohort study.

For each report type, one clinical note was randomly selected for each patient. Therefore, the dataset contains 100 notes for each report type. Taking each Chinese character as one token and each English word as one token, 203, 883 tokens are found in the dataset. The average length of sentences in terms of tokens is 23, with a minimum length of 2 and a maximum length of 447. The average numbers of sentences per report type are listed in Table 1.

To determine the annotation scope, an information model that contains a rich set of fine-granular genotypic and phenotypic information was designed by domain experts. In total, 41 entities (i.e., clinical concepts and their attributes) and 34 relations between them were defined (Ref. Supplementary Table 1), covering demographic, medical history, family history, pathology test, radiology test and Immunohistochemical test information. The total numbers of entities of each semantic type are listed in Table 2. An illustration of the specific information that is extracted for breast cancer patients is presented in Fig. 1. Two students with medical background were trained and annotated the dataset. The overall inter-annotator agreement was 0.81 for entities and 0.89 for relations, as measured by Cohen's Kappa.

2.2. Pretraining of the clinical language model

The language model of Chinese clinical texts is generated using BERT to pretrain the 3GB corpus of clinical texts. At present, the model generation has taken 300,000 training steps. Specifically, it is based on the original language model of Chinese wiki data that was pretrained in BERT. The original language model was pretrained and transferred to the medical domain.

2.3. Information extraction for breast cancer

Our system consists of three components: preprocessing, NER and relation extraction. The details of the system components are described below.

2.4. Preprocessing

The preprocessing module includes the basic steps of sentence

Table 1
Average number of sentences in each type of clinical note. For discharge summaries, redundant information from other report types is removed and only the "Procedure of Diagnosis and Treatment" text is kept.

Clinical Report type	Average number of sentences
Encounter notes	66
Operation records	16
Pathology notes	5
Radiology notes	6
progress notes	5
discharge summaries	3

Table 2
Total number of entities for each semantic type in breast cancer notes.

Entity type	Number	Entity type	Number
Original site	516	ERβ	421
Disorder	6714	The specific distance between the tumor and the nipple	720
Ultrasonography of axillary lymph nodes	837	P53	624
Characteristics	1415	Size (Pathology)	650
Negation	3882	Nipple	455
Temporal expression	2665	Ki67	710
Test value	6411	Number of children	693
Smoking	897	CK5/6	653
Surgery	1475	EGFR	308
Body location	3184	P53	624
Bleeding flow	1548	Pathology test	600
Pathological type	1106	ERα	571
Her-2	733	Regional Lymph nodes	984
Differentiation grade	771	Alcohol	549
PR	733	Echo	419
Tumor Volume	1134	Test units	741
Abnormality	2323	Menarche age	511
Form	1091	Menstrual days	511
Boundary	973	Menstrual cycle	509
Tumor direction	895	Age of amenorrhea	295
Radiology test	837	Total	51,688

boundary detection and tokenization. However, the sentence boundaries are not well formed in clinical text. For example, there are fragments in the text without any punctuation, and some notes only use commas throughout whole paragraphs without any periods. Therefore, we used spaces, commas and the length of clauses to split the text into sentences instead of only using the periods. The default length of a sentence in BERT is 128. If the input sentence is longer than 128, the extra length of the sentence will be automatically deleted. As mentioned above, some sentences in clinical text were much longer, reaching maximum length of 447. To maintain the complete information of sentences, they were split using spaces or commas once the length exceeded 128 tokens. If the length of the split clauses were still longer than 128, they would be further split until all the segmented sentences could satisfy the length constraint of BERT.

2.4.1. Input representation of named entity recognition

For NER, annotated data were transformed into the BIO format, where "B" represents the beginning of an entity, "I" represents other words inside an entity, and "O" represents all other nonentity words. The NER models will predict the BIO labels for the input sentences, which will be transformed back to named entities.

2.4.2. Fine tuning the BERT model for named entity recognition

Among different deep learning approaches, the BI-LSTM-CRF is widely used and achieves good performances in different NER tasks [14–17]. The BI-LSTM-CRF uses a BI-LSTM to score all possible labels for each token in a sequence, and it predicts a token's label using its neighbor's information in a CRF layer [18,19]. The output vectors of the pretrained language model using BERT are used as the features of the input layer of the BI-LSTM-CRF. Both the BI-LSTM-CRF network and the BERT network are tuned during the training process (Fig. 2).

Given the annotated entities in sentences, the relation extraction task can be transformed into a classification problem. A classifier can be built to determine the categories of all possible candidate relation pairs (e_1, e_2), where entities e_1 and e_2 are from the same sentence. We generated candidate pairs by pairing each concept and another entity with a semantic type that matches an attribute of the concept.

2.4.3. Input representation for relation extraction

To represent a candidate relation pair in an input sentence, we used the semantic type of an entity to replace the entity itself. Semantic types refer to the entity types that are defined in our information model, such as disease, medicine and temporal expressions. The mentions of entities are directly replaced by their semantic types in the sentences. For example, there are many mentions of "cyclophosphamide" in the text, and they were directly replaced by "Medicine".

2.4.4. Fine tuning the BERT model for relation extraction

Devlin et al.'s BERT model [12] was used, and a linear classification layer was added on top to predict the label of a candidate pair in a sentential context (Fig. 3). In detail, BERT adds a classification token [CLS] at the beginning of a sentence input, and the output vector was used for classification. As is typical with BERT, we used a [CLS] vector as the input to the linear layer for classification. Then, a softmax layer was added to output labels for the sentence.

2.5. Evaluation

The primary evaluation metric is the lenient micro F1 score (Eq. 3). We used 10-fold cross-validation on the training dataset to optimize the parameters for the models.

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (1)$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (2)$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

Moreover, the pairwise *t*-test was conducted using the model outputs of our proposed method and each of the other implemented methods in order to check the statistical significance of the performance differences.

2.6. Implemented baselines for NER

1 The CRF is known for its good NER performance, including that in medication information extraction [20–22]. Therefore, we chose it as a strong baseline for comparisons with deep learning approaches. Common NER features are used, including the n-gram, prefix-suffix, orthographic, discretized word embeddings, etc.
2 The original BI-LSTM-CRF model without fine tuning with BERT is also used as a baseline. The Chinese character embedding was trained on a corpus of 3 Gigabytes of clinical notes [23]. The word2vec package was used to generate the character embeddings with a dimension size of 300 [23]. Here, we revised the architecture of the BI-LSTM-CRF from Lample et al.'s work for Chinese NER [18]. Chinese character embeddings are used as the input embeddings instead of the word embeddings and the character embeddings within each word for English text.
3 A simple keyword-based search baseline was implemented by using the entity mentions in the training dataset with frequencies higher than 5 as the lexicon and exact matches to find the entities in the test dataset.

2.7. Implemented baselines for relations

1 Support Vector Machine (SVM): The SVM has been used in previous relation classification tasks for clinical text and achieved good performance [9,11,13]. We used the SVM as a baseline method to compare it with other deep learning methods in the end-to-end and relation classification tasks.
2 The method that was proposed in Zhou et al. [13] which used biGRU with character and sentence attention for Chinese relation extraction, was used as another baseline. Unlike BERT, attention-based Bi-

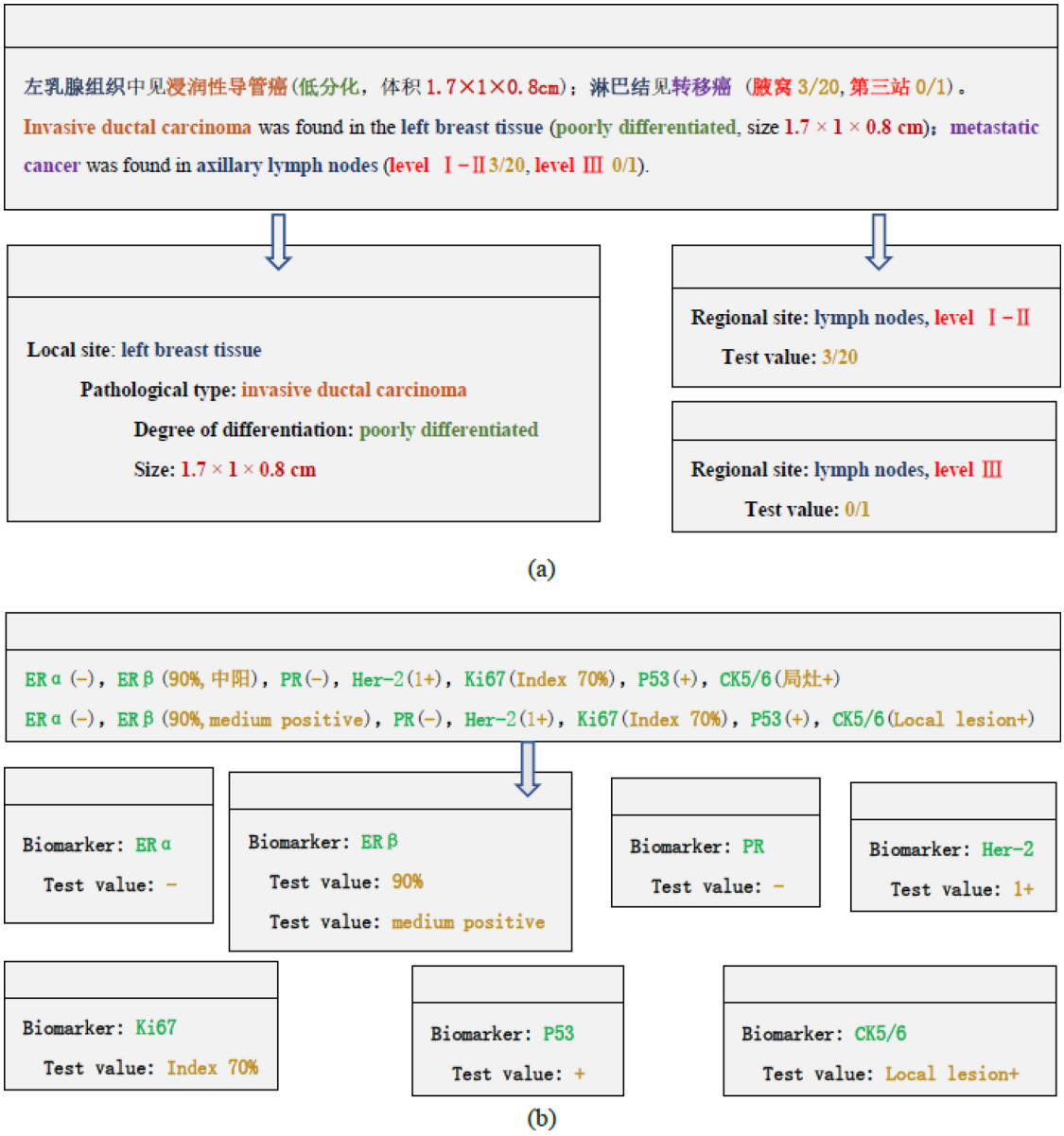

Fig. 1. Illustration of the breast cancer information model. The original Chinese clinical text and its English translation are presented. The extracted clinical concepts and their attributes are also listed, and they are highlighted in different colors. (a) Pathology test. (b) Immunohistochemical test.

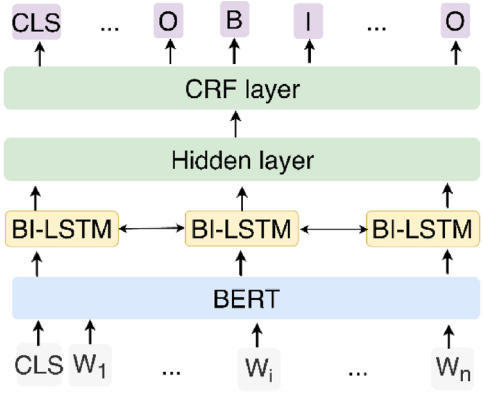

Fig. 2. The architecture of fine-tuning BERT for named entity recognition based on the Bi-LSTM-CRF. CLS represents the [CLS] token and Wi represents the tokens in a sentence.

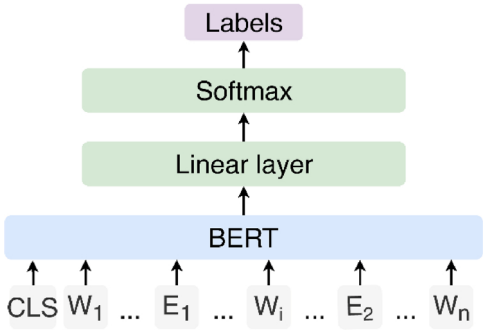

Fig. 3. The architecture of fine-tuning BERT for relation extraction. CLS represents the [CLS] token and Wi represents the tokens in a sentence. Ei represents the entities in a sentence.

LSTM was used to capture the most important semantic information from characters and sentences. The semantic information of sentences and characters was then used for relation extraction. In

addition, the input features were the character embeddings that were generated from word2vec instead of a pretrained language model, as in BERT.

3 The method in Li et al. [24]used the dependency structures of sentences to enhance the relation extraction performance. A convolution was used to represent the dependency and Bi-LSTM was used to model other syntactic/semantic information.

The parameters and hyperparameters that are used in the optimal model of this study are as follows.

Both the input and hidden layers were set to 200 neurons. The dropout rate of each layer is 0.5, and the batch size is set to 32. The L2 regularization was only applied to the input layer with 0.0001 as the coefficient. The hidden layer used tanh as the activation function. The logistic sigmoid function was used in the last layer. The overfitting of the model was controlled by using regulation strategies, including using L2 regulations, dropout with a probability of 50% and early stopping when the loss has no significant changes.

3. Results

3.1. Named entity recognition

Overall, the NLP-based information extraction system achieved a Precision of 0.927, a Recall of 0.939 and an F-measure of 0.935 (0.700–1.000) for entity recognition (Table 3). The performance range when individually evaluating each semantic type is 0.700–1.000. The lowest and highest F-measures are 0.700 and 1.000, respectively. The detailed performance of the complete list of entity types can be found in Supplementary Table 2. The BI-LSTM-CRF model significantly enhanced the recall of the CRF, which was further improved by using the language model of the pretrained BERT for the input features. The overall performance of the keyword-match baseline method is extremely low. One potential reason is that there are many variants in the attributes, such as in the test values of biomarkers, making it inappropriate to use the keyword/lexicon-based method for NER.

To provide a detailed illustration, the performances of some clinical concepts and their attributes are listed in Table 4.

3.2. Relation classification

The system that is built on fining-tuning BERT achieved a Precision of 0.976, a Recall of 0.959, and an F-measure of 0.967 (0.500˜1.000) for relation recognition (Table 5). The performance range when individually evaluating each relation type is 0.500˜1.000. The lowest and highest F-measure are 0.500 and 1.000, respectively. The detailed performance of the complete list of relation types can be found in Supplementary Table 2. In fact, the SVM already acts as a strong baseline with an F-measure of 0.936, and the deep learning models further enhanced the performance. For a detailed illustration, the performances of some relations between clinical concepts and their attributes are listed in Table 6. The relation extraction models were trained and tested using the gold-standard NE annotations. To assess the practical performance of the system in the real environment, we also evaluated the end-to-end performance by using automatically generated NEs as the relation extraction inputs, and the performances were promising, reaching a precision of 0.921, a recall of 0.820, and an F1 of 0.868.

Table 3
Overall clinical concept extraction performances from clinical breast cancer notes.

Method	Precision	Recall	F1
BERT + Bi-LSTM-CRF	0.927	0.939	0.935*
Bi-LSTM-CRF	0.883	0.892	0.887
CRF	0.855	0.819	0.837
Keyword-match	0.359	0.743	0.484

* Means the performance difference is statistically significant (p-value < 0.05).

Table 4
Extraction performances of examples of clinical concepts from clinical breast cancer notes using our proposed methods.

Concept	Precision	Recall	F1
Tumor direction	1	0.8	0.889
Biomarker	1	1	1
Surgery	0.957	0.957	0.957
Temporal information	0.895	0.971	0.932
Regional lymph node	0.9	1	0.947
Disorder	0.974	0.991	0.982
Ultrasonography	0.846	0.917	0.88
Pathological type	0.8	0.923	0.857
Abnormality	0.889	0.8	0.842
Body location	0.786	0.786	0.786

Table 5
Overall concept relation extraction performances from clinical breast cancer notes.

Method	Precision	Recall	F1
Fining-tuning BERT	0.976	0.959	0.967*
Dependency-path	0.951	0.943	0.945
Attention + BiLSTM	0.954	0.947	0.951
SVM	0.942	0.929	0.936

* Means the performance difference is statistically significant (p-value < 0.05).

Table 6
Performance of the relations that were extracted between clinical concepts and their attributes for breast cancer using our proposed methods.

Concept	Attribute	Precision	Recall	F1
Smoking	Negation	1	1	1
Biomarker	Test value	0.983	1	0.991
Abnormality	Negation	0.989	0.994	0.992
Blood flow	Body location	0.931	0.964	0.947
Body location	Size(Pathology)	0.966	1	0.983
Disorder	Temporal expression	0.967	0.935	0.951
Abnormality	Body location	0.991	0.913	0.95
Pathological type	Degree of histological differentiation	1	0.988	0.994
Immunohistochemical test	Temporal expression	0.896	0.892	0.876
Disorder	Relative	0.6	1	0.75

4. Discussion

With the continuous accumulation of data, clinical data management has become a difficult task. Especially, it is even more complex for Chinese clinical text, which is a mixture of terminologies in Chinese, English (e.g., brca1, Cerb-2) and numeric information. NLP models that were designed for general language understanding are usually ineffective in clinical text processing tasks [25]. Meanwhile, it is very expensive and time-consuming to construct a large-scale training data set. Therefore, deep learning methods, which use contextual embeddings that are learned in an unsupervised way from large-scale unlabeled clinical text as input features, usually outperform other state-of-the-art machine learning methods on clinical NLP tasks.

Breast cancer is the most common malignant tumor among women, and its incidence rate is increasing in China. It needs the integration of multiple clinical variants in EHRs for accurate diagnosis and treatment. This study takes the initiative and designs a breast cancer information model representing a comprehensive collection of clinical concepts and

their relations. One advantage of this study is that a comprehensive list of clinical variables that were related to breast cancer were extracted based on a predefined information model. The model included basic demographical information on hospitalization records (e.g., age, height, weight, marital status), lifestyle (e.g., smoking, drinking), and epidemiological susceptibility factors (e.g., menstruation, childbearing, breast-feeding); medical complications; preoperative imaging examinations; surgical records; postoperative hospital processes; and postoperative pathology information. It is a whole-course information representation that can provide useful information for breast cancer clinical treatments and research.

Moreover, the latest deep learning language model BERT [12] was used for the information extraction from Chinese clinical breast cancer notes. BERT demonstrated its superiority over other state-of-the-art deep learning methods and traditional feature-engineering-based machine learning methods on multiple NLP tasks such as NER and sentence classification [12,13]. BERT defined a novel language modeling framework by using bidirectional transformer architectures. It more effectively represents the contextual information in large-scale unlabeled text, which can be conveniently transferred to downstream NLP tasks. In this study, a clinical language model is first pretrained using BERT on an unlabeled corpus of Chinese clinical text. The linguistic knowledge that was embedded in the pretrained clinical language model was then applied to our IE tasks. For NER, the context embeddings that were pretrained using BERT were used as the input features of the Bi-LSTM-CRF model and were fine-tuned using the annotated breast cancer notes. Bi-LSTM-CRF is widely used around the globe for NER due to its strength in combining the Bi-LSTM architecture for context modeling and a CRF layer for further sequence optimization. Augmented with the language model that was pretrained using BERT, the model achieved a Precision of 0.927, a Recall of 0.939 and an F-measure of 0.935. In addition, we proposed an approach to fine-tune BERT for relation extraction. The relation extraction task was considered as a classification task where the two entity mentions in the input sentence were replaced with their semantic types. Such a transformation dramatically generalized the contextual information and achieved a Precision of 0.976, a Recall of 0.959, and an F-measure of 0.967. The experiments showed that our system outperformed other benchmark methods for both NER and relation extraction (Tables 1, 3).

Our clinical IE system can automatically identify and record two important indicators of biomarkers, their intensities (e.g., +, ++, and ++++) and proportions (e.g., 95%). For complex clinical variables including the complex locations of tumors, how far away a tumor is from the nipple, and whether the blood flow is rich or not, they are also well recognized via deep learning methods.

Currently, the majority of the works on clinical information extraction have been conducted for and reported in English [4–9]. However, it is essentially important to develop high-performance systems for the clinical information extraction of other languages including Chinese to promote the secondary use of EHRs for different populations worldwide. Clinical Chinese and English texts have different characteristics in terms of the linguistic traits and the writing styles of physicians. For example, Chinese use each character as a token, while the token in English is usually a word. In addition, Chinese does not have the explicit passive tense as in English, which needs to be detected using relations with temporal expressions. Even though the findings in different languages cannot be directly compared, our models produced promising results that are on par with English models. For example, an F-measure of 0.857 was achieved for the recognition of the pathological type, which was comparable to the findings in Yala et al. [6].

Limitations and future works: One limitation of our current work is that the clinical notes of only 100 patients are employed for information extraction, which may not comprehensively cover more diverse clinical concepts and attributes. Several rare concepts and relations achieved relatively low performance. Additional data will be annotated to increase the size of the training set in the next step. In our experiments, the evaluation used 10-fold cross validation due to the sparsity of certain entity types and relations in the current data set. In our future work, the data set will be expanded, which will be split into a training, validation and test set for the evaluation. Despite the limited data scale, our system achieved promising performance for this practical clinical application. In the future, extracted information will be normalized to standard terminologies such as the ICD10 for practical clinical applications. Two other potential future extensions of this system include the following. (1) It will be integrated with image data processing modules for CT, MR imaging and pathological microscopic images. Combining image analysis and text processing will provide better findings from radiological tests for practical clinical applications. (2) Due to the high accuracy of the extracted information, statistical association analysis will be conducted to answer questions about etiology, epidemiology and the prognostic factors. For example, these questions may include the following. "What risk factors may be related to breast cancer?" "What factors affect the prognosis of breast cancer, such as lymph node status, estrogen levels and progesterone receptor status?"

5. Conclusion

In summary, this study takes the initiative and designs a breast cancer information model that comprehensively covers clinical concepts and relations. In addition, a clinical language model based on BERT is pretrained and fine-tuned for information extraction from multiple types of clinical notes. Experiments demonstrate that our proposed method achieves promising results, which outperform other state-of-the-art algorithms. The deep learning-based clinical IE systems that are built in this study could be applied to facilitate the large-scale secondary use of EHRs for clinical and translational breast cancer research such as precision medicine and clinical outcome predictive modeling. The findings in this study could also assist in establishing high-performance clinical IE systems in other medical domains and languages.

Summary table

What was already known on the topic

- The clinical information of breast cancer patients is abundant in multiple types of clinical notes.
- Natural langue processing methods can effectively extract information from clinical text.
- Current works in English are mainly focused on extracting breast cancer information from pathology notes and radiology notes.
- Few works have been conducted to extract information from clinical notes in Chinese.

What this study added to our knowledge

- A novel information model is designed to represent the whole-course patient information of breast cancer.
- This is the first study to extract comprehensive clinical information from multiple types of breast cancer notes.
- The information extraction models that are generated by fine-tuning BERT outperform current state-of-the-art methods.

Author contributions

Designed the information model: XZ, TQ, JM, and QS. Conceived and designed the experiments: XZ, and YZ. Performed the experiments: YZ, QZ, and YR. Analyzed the data: XZ, YZ, QZ, and YR. Wrote the paper: XZ, and YZ. Revised the manuscript: TQ, JM, and QS.

Funding

The research was supported by the Chinese Academy of Medical Science Initiative for Innovative Medicine (2017-I2M-2-003) and the Chinese National Key R&D (2018YFC0116901).

Declaration of Competing Interest

None.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.
[2] R.L. Segal, K.D. Miller, A. Jemal, Cancer statistics, 2018, CA Cancer J. Clin. 68 (2018) 7–30.
[3] W. Chen, K. Sun, R. Zheng, H. Zeng, S. Zhang, C. Xia, et al., Cancer incidence and mortality in China, 2014, Chin. J. Cancer Res. 30 (1) (2018) 1.
[4] J.M. Buckley, S.B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A.K. Belli, et al., The feasibility of using natural language processing to extract clinical information from breast pathology reports, J. Pathol. Inform. (June) (2012) 30 [Internet].
[5] A.E. Wieneke, E.J.A. Bowles, D. Cronkite, K.J. Wernli, H. Gao, D. Carrell, et al., Validation of natural language processing to extract breast cancer pathology procedures and results, J. Pathol. Inform. (June) (2015) 23 [Internet].
[6] A. Yala, R. Barzilay, L. Salama, M. Griffin, G. Sollender, A. Bardia, et al., Using machine learning to parse breast pathology reports, Breast Cancer Res. Treat. 161 (January (2)) (2017) 203–211.
[7] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, D. Page, Information extraction for clinical data mining: a mammography case study, 2009 IEEE International Conference on Data Mining Workshops, (2009), pp. 37–42.
[8] D.A. Sippo, G.I. Warden, K.P. Andriole, R. Lacson, I. Ikuta, R.L. Birdwell, et al., Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing, J. Digit. Imaging 26 (October (5)) (2013) 989–994.
[9] T.A. Patel, M. Puppala, R.O. Ogunti, J.E. Ensor, T. He, J.B. Shewale, et al., Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods, Cancer 123 (1) (2017) 114–121.
[10] J.X. Qiu, H. Yoon, P.A. Fearn, G.D. Tourassi, Deep learning for automated extraction of primary sites from cancer pathology reports, IEEE J. Biomed. Health Inform. 22 (January (1)) (2018) 244–251.
[11] S. Miao, T. Xu, Y. Wu, H. Xie, J. Wang, S. Jing, et al., Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches, Int. J. Med. Inform. 119 (2018) 17–21.
[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert, Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv Preprint arXiv:181004805, (2018).
[13] Yuqi Si, Jingqi Wang, H. Xu, K.E. Roberts, Enhancing Clinical Concept Extraction with Contextual Embedding, (2019).
[14] A.N. Jagannatha, H. Yu, Structured prediction models for RNN based sequence labeling in clinical text, Proceedings of the Conference on Empirical Methods in Natural Language Processing Conference on Empirical Methods in Natural Language Processing, (2016), pp. 856–865 November; 2016.
[15] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, et al., Entity recognition from clinical texts via recurrent neural network, BMC Med. Inform. Decis. Mak. 17 (July (S2)) (2017) 67.
[16] Z. Liu, B. Tang, X. Wang, Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, J. Biomed. Inform. 1 (November (75)) (2017) 34–42.
[17] B. Dandala, V. Joopudi, M. Devarakonda, F. Liu, A. Jagannatha, H. Yu, IBM research system at MADE 2018: detecting adverse drug events from electronic health records, Proceedings of Machine Learning Research, (2018), pp. 39–47.
[18] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, Proceedings of NAACL-HLT (2016) 260–270.
[19] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, CoRR (August (9)) (2015) abs/1508.01991.
[20] J. Patrick, M. Li, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, J. Am. Med. Inform. Assoc. 17 (September (5)) (2010) 524–527.
[21] A.B. Chapman, K.S. Peterson, P.R. Alba, S.L. DuVall, O.V. Patterson, Detecting adverse drug events with rapidly trained classification models, Drug Saf. (January (16)) (2019) 1–10.
[22] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, et al., Extraction of adverse drug effects from clinical records, Stud. Health Technol. Inform. 160 (Pt 1) (2010) 739–743.
[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems (2013) 3111–3119.
[24] L. Zhiheng, X. Jun, X. Yang, W. Qiang, Z. Yaoyun, X. Hua, Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text, BMC Medical Informatics and Decision Making, (2018).
[25] M. Habibi, L. Weber, M. Neves, D.L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics 33 (14) (2017) 37–48.