Received April 29, 2019, accepted May 14, 2019, date of publication May 20, 2019, date of current version June 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917922

An Annotation Model on End-to-End Chest Radiology Reports

XIN HUANG[1,2], **YU FANG**[1], **MINGMING LU**[1], **YAO YAO**[1], **AND MAOZHEN LI**[3]

[1]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[2]Software College, Jiangxi Agricultural University, Nanchang 3300029, China
[3]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, U.K.

Corresponding author: Yu Fang (fangyu@tongji.edu.cn)

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 16511102800, and in part by the Fundamental Research Funds for the Central Universities under Grant 22120180117.

ABSTRACT Annotating radiographic images with tags is an indispensable preliminary work in computer-aided medical research, which requires professional physician participated in and is quite time-consuming. Therefore, how to automatically annotate radiographic images has become the focus of researchers. However, image report texts, containing crucial radiologic information, have not to be given enough attention for images annotation. In this paper, we propose a neural sequence-to-sequence annotation model. Especially, in the decoding phase, a probability is first learned to copy existing words from report texts or generate new words. Second, to incorporate the patient's background information, ''indication'' section of the report is encoded as a sentence embedding, and concatenated with the decoder neural unit input. What's more, we devise a more reasonable evaluation metric for this annotation task, aiming at assessing the importance of different words. On the Open-i dataset, our model outperforms existing non-neural and neural baselines under the BLEU-4 metrics. To our best knowledge, we are the first to use sequence-to-sequence model for radiographic image annotation.

INDEX TERMS Annotation, chest radiology report, deep learning, end-to-end model, indication.

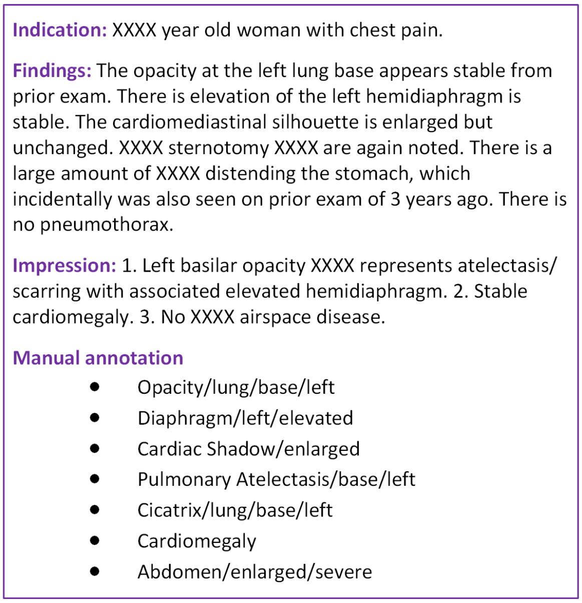
I. INTRODUCTION

The two essential elements of computer-aided medical research consist of sufficient data and high-quality labels. In fact, hospital never lacks data because a massive amount of data was stored in Electronic Health Record (EHR) and Picture Archiving and Communication Systems (PACS). Therefore in many tasks, extracting the correct labels from a large amount of data through automated tools goes first [1]. For instance, mapping radiographic images to pathology requires the automatic annotation tools to extract the correct pathology labels from the radiology reports. These labels are critical for subsequent tasks such as radiographic image retrieval, disease population analysis, and clinical behavior analysis. Low-accuracy pathology extraction tools may result in a high percentage of false labels in large radiology corpora, leading to bias in subsequent research results.

In general, the annotation work is divided into three major tasks as follows: the first one is to extract the pathological terms, that is, to exclude the non-disease-related terms in the report and to indicate the position of the pathological terms in the report. Because different doctors often use different terms when describing the same disease or condition, the second task is to unify the description of the same pathology term. The third task is map the relationship between disease and disease description, because the description of the disease shows more details about the disease, such as location, degree of illness, etc. Unlike most private dataset-based researches, we believe that the research on public datasets offers a contrast with existing effective methods and benchmark test for future progress. Therefore, we choose the public open-i[1] dataset as our research object. As shown in Fig. 1, the finding and impression sections of the radiology report convey important information about the radiological image. The indication section contains key background information such as patient gender, age, and past medical history. The Manual annotation section is labeled by a professional radiologist, and each annotation consists of one or more medical terms. In fact, some of the medical terms included in the Manual annotation are copied directly from the report, such as: *''opacity''*,

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

[1] https://openi.nlm.nih.gov/gridquery.php?q=&it=xg&coll=cxr



Indication: XXXX year old woman with chest pain.

Findings: The opacity at the left lung base appears stable from prior exam. There is elevation of the left hemidiaphragm is stable. The cardiomediastinal silhouette is enlarged but unchanged. XXXX sternotomy XXXX are again noted. There is a large amount of XXXX distending the stomach, which incidentally was also seen on prior exam of 3 years ago. There is no pneumothorax.

Impression: 1. Left basilar opacity XXXX represents atelectasis/scarring with associated elevated hemidiaphragm. 2. Stable cardiomegaly. 3. No XXXX airspace disease.

Manual annotation

- Opacity/lung/base/left
- Diaphragm/left/elevated
- Cardiac Shadow/enlarged
- Pulmonary Atelectasis/base/left
- Cicatrix/lung/base/left
- Cardiomegaly
- Abdomen/enlarged/severe

FIGURE 1. **A sample of radiology report in openi dataset.**

"cadiomegaly", etc. (task one); some terms are not found in the report, for example: *"Diaphragm"* is actually a report Normalization of *"hemidiaphragm"* (task two). At the same time, *"Pulmonary Atelectasis/base/left"* embodies the relationship between disease and location (Task three).

For the time being, these three tasks are usually handled separately. We believe that can learn from the good experience gained from deep learning in natural language and combine the three tasks. We consider the annotations containing disease and disease descriptions as a special text summary of a generic NLM indexing guide, and propose an end-to-end automatic labeling model. After training, our model can decide by probability whether to choose to copy the existing words in the radiology text as a label or to generate a new word as a label. At the same time, considering that the radiologist will also pay attention to the patient background information when the diagnosis is made, although the indication part does not directly contain the medical vocabulary required for labeling, we believe that there is a reasoning relationship between them. Therefore, our model adds processing branches for indication information representation. The results show that the addition of indication information significantly improves the performance of the model. Our work has three main contributions:

- We propose a neural sequence-to-sequence annotation model. Specially, in the decoding phase, a probability is first learned to copy existing words from report texts or generate new words.
- We propose a new customized annotation model to this task that improves over existing methods by better leveraging study "indication" information.
- We devise a more reasonable evaluation metric for this annotation task, aiming at assessing the importance of different words.

The rest of this paper is organized as follows. Section II reviews the related work for natural language inference. Section III details the design of the proposed model. Section IV and V present and discuss the experimental settings and results, respectively. Finally, we draw conclusion in Section VI.

II. RELATED WORK

A. ANNOTATION FROM MEDICAL TEXT

The NLM Medical Text Indexer [2] provided an automatic indexing of medical literature. Gobbel *et al.* [3] developed a report annotation model called rapt in the study of care quality during the treatment of heart failure, and the annotation accuracy was improved by the mechanism of iteration. Tonin [4] developed a machine learning based annotation tool, which can extract the mentions or indications for coronary artery disease in unstructured clinical reports. Demner-Fushman *et al.* [5] presented a small number of chest radiology terms. Zhou *et al.* [6] implemented the relationship extraction method based on a semi-supervised bootstrapping framework. Mostafiz and Ashraf [7] trained a DNN-based Named Entity Recognition (NER) model to extract the key concept words from radiology reports, which results demonstrate the inadequacy of generic APIs for pathology extraction task and establishes the importance of domain specific model training for improved results. Wang *et al.* [1] mined the disease terminology in the report by using DNorm [8] and MetaMap [9] tools and released a large chest dataset ChestX-14. Irvin *et al.* [10] developed an automated rule-based labeler to extract observations from the free text radiology reports to be used as structured labels for the images, which set up in three distinct stages: mention extraction, mention classification, and mention aggregation. Banerjee *et al.* [11] proposed an unsupervised hybrid method-Intelligent Word Embedding (IWE) that combines neural embedding method with a semantic dictionary mapping technique for creating a dense vector representation of unstructured radiology reports. Most of the previous studies on chest radiographs were focused on the extraction of disease, ignoring the location and severity of the disease.

B. NEURAL SUMMARIZATION MODELS

Traditional machine learning only extracts keywords from the text to generate a summary. In contrast, the summary model based on the neural network model supports the generation of summaries with new words and phrases. Rush *et al.* [12] first applied attention-based neural encoders and neuro-language model decoders to this task. Nallapati *et al.* [13], a cyclic neural network, based on the encoder-decoder model was used to process the task. In order to solve the problem that the neural model based on the fixed vocabulary cannot handle the unknown vocabulary generation, Merity *et al.* [14] proposed a pointer-generator model by duplicating the attention mechanism of input text elements. What's more, See *et al.* [15] further proposed a coverage mechanism to address the repetitive problem in generating summaries. In a recent study, Paulus *et al.* [16] applied intensive learning to the generation of summaries standing on the work of the predecessors. Moreover, Chen and Bansal [17] adopted a new idea and proposed a model that selects sentences first and then rewrites sentences, which achieved better results.

Most of the previous studies on chest radiographs were focused on the extraction of disease, ignoring the location and severity of the disease. Although Shin *et al.* [18] tried to generate a complete annotation containing the disease and the description of the disease, his job was to automatically generate annotations from the image features, and the results were not good. In addition, the general deep natural language model usually does not consider medical attributes and is not fully suitable for medical report research. These shortcomings are the goal of our research.

III. METHODS

A. OVERVIEW

At a high level, our approach is to use the encoder-decoder architecture to implement the task. The encoder accepts a sequence as input and encodes the information in the sequence as a hidden state representation; the decoder then decodes the input representation into an output sequence. For the Findings and Impression sections in a given radiology report text, $X = \{x_1, x_2, \ldots, x_N\}$ where N is the length of the text. The goal of model is to generate a corresponding label Y based on X. $Y = \{y_{11}, y_{12}, \ldots, y_{31}, y_{32}, \ldots y_{KL}\}$, where K is the number of labels, and L is the length of the label.

B. SEQUENCE-TO-SEQUENCE ATTENTION MODEL

In the work,the encoder uses a Bi-directional Long Short-Term Memory (Bi-LSTM) network to proceed. The Findings and Impression sections from the report are combined as the input; the merged sequence is represented as $X = \{x_1, x_2, x_3, \ldots, x_N\}$. x is encoded into hidden state vectors with:

$$h = BiLSTM(x) \tag{1}$$

where $h = \{h_1, h_2, h_3, \ldots, h_N\}$.Specially, h_N is the result of the bidirectional last hidden states. The output of the decoder is a sequence that is standardized by MeSH vocabulary and arranged in a specific format. This paper uses the unidirectional Long Short-Term Memory (LSTM) network in the decoder section, whose initial state s_0 is the output h_N of the encoder. On the step t,the decoder receives the previous decoder state s_{t-1} and the previous generated token y_{t-1},the decoder current state s_t calculated as:

$$s_t = LSTM(s_{t-1}, y_{t-1}) \tag{2}$$

This method only uses s_t to connect the encoder and decoder. Therefore, the encoder needs to compress the entire sequence information into a fixed-length vector, which brings many limitations. As the length of the input sequence increases, the information input first is diluted by the information input later. For better decoding, we takes the attention mechanism [19], [20] to instruct the decoder to generate the next word according to the probability distribution of the source word. Attention distribution a^t can be calculated by s_t and h_i.

$$e_i^t = v^T \tanh(w_h h_i + w_s s_t) \tag{3}$$

$$a^t = softmax(e^t) \tag{4}$$

where v, W_h, W_s are parameters that need to be learnable. a^t is used to calculate the context vector h_t^*:

$$h_t^* = \sum_i a^t h_i \tag{5}$$

h_t^* contains the important information of the decoding. We finally get the probability distribution P_{vocab} of the output word through h_t^*:

$$P_{vocab}(y_t|x, y_{<t}) = softmax(V' \tanh(V[s_t; h_t^*])) \tag{6}$$

where V',V are parameters that need to be learnable.

C. COPY MECHANISM

While the Sequence-to-Sequence attention model can generate annotations from a given MeSH vocabulary, in many cases, the annotation vocabulary is obtained directly from the Findings and Impression sections. Obviously, it is more efficient to copy the words that need to be annotated directly from the input text, so a pointer-generator network that is similar to the one described in [15] is added to the model. In the process of decoding, the model is allowed to generate a word from the MeSH vocabulary by generating the probability p_{gen}, or duplicating a word directly from the sentence with a probability of $1-p_{gen}$. The p_{gen} calculation method of the model is as follows:

$$p_{gen} = \sigma(w_{h^*}^T + w_s^T s_t + w_y y_{t-1}) \tag{7}$$

where σ is the sigmoid function and $y_{(t-1)}$ is the output of the previous decoder. w_{h^*}, w_s, w_y are parameters that need to be learnable. At last, the probability distribution P of the Pointer-Generator Network output is obtained:

$$P_{vocab}(y_t|x, y_{<t}) = p_{gen} P_{vocab}(y_t) + (1 - p_{gen}) \sum_{i:w=y_c} a^t \tag{8}$$

In the report, the same word often appears multiple times, for example, when the left lung is blurred and the left corner is raised, the “left” word repeat twice. So we made some changes, the “coverage” mechanism of our model is prohibited.

D. REPRESENTATION OF INDICATION INFORMATION

The information in the Indication section of the radiology report is critical to diagnosis, since background information such as patient’s age, gender, physical condition, and discomfort location is only mentioned in Indication. Additionally, it is not well performed to generate the final label by combining Indication with Findings and Impression into a single piece of information as input to the our model, because

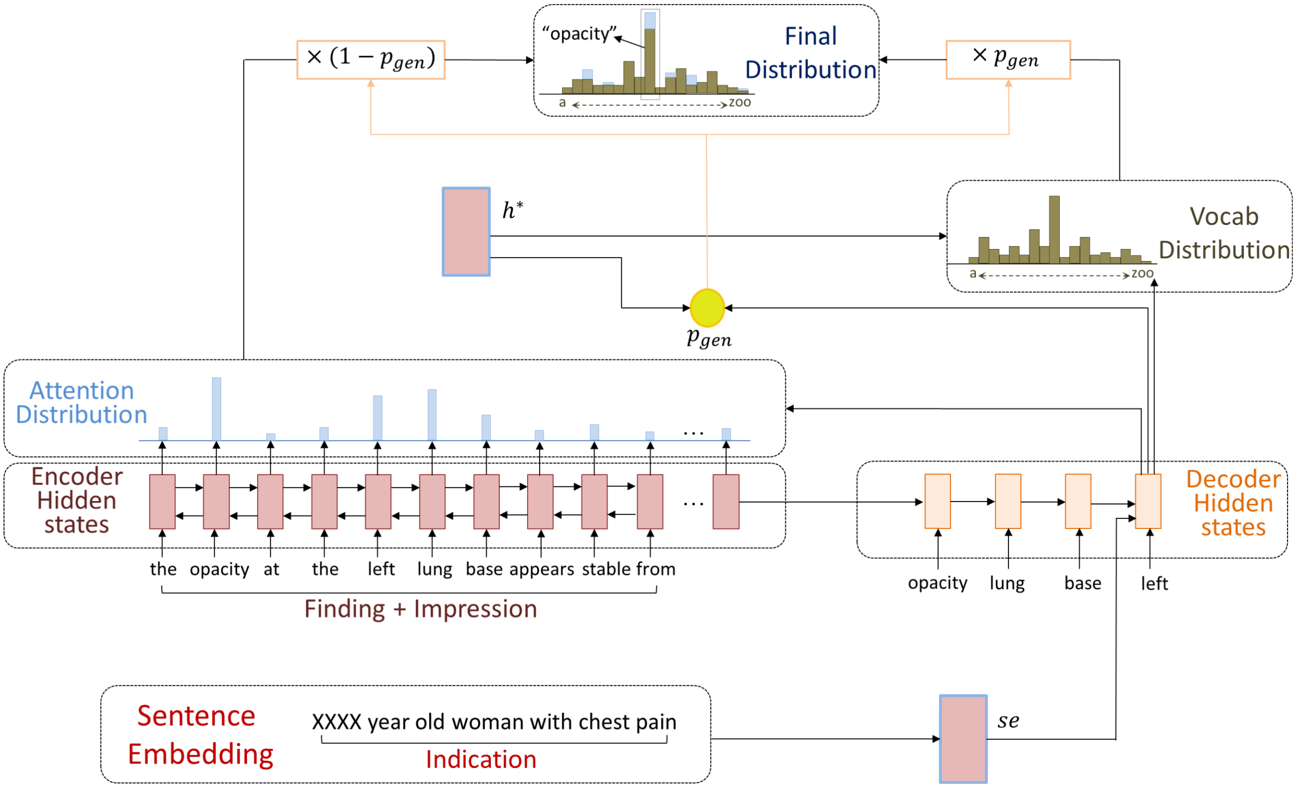


FIGURE 2. Annotation model architecture.

the information contained in the indication has an inferential relationship with Findings and Impression. Further, direct merging can result in insufficient modeling due to overly complex textual information, which in turn affects the quality of the final annotation. In order to solve this problem, proposes to use the information contained in the indication to guide the decoding separately. Fig. 2 shows the architecture of our model.

We use Sentence Embedding [21] to represent an indication. Each indication is mapped to a unique vector and represented by a column of matrix S. The column index of S is the serial number of the indication in the document, and the Sentence Embedding se is calculated as follows:

$$se = b + Uh(s_{t-k}, \ldots, s_{t+k}; S) \tag{9}$$

where U, b is the softmax parameter. h is constructed by average of sentence vectors extracted from S. Finally, we modify the kernel of LSTM. We add Sentence Embedding e to the process of decoding. At the same time, in order to maximize the guiding effect of e, we remove the bias trems (formula 10) in LSTM, which is calculated as follows:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ u_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot \begin{bmatrix} s_{t-1} \\ y_{t-1} \\ se \end{bmatrix} \tag{10}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot u_t \tag{11}$$

$$s_t = o_t \cdot \tanh(c_t) \tag{12}$$

where i_t, f_t, o_t are input gates, forgetting gates and output gates respectively. c_t is LSTM's internal cell. W is the weight matrix. · refers to element multiplication. The rest of the model, including the attention, vocabulary distribution, and final distribution calculations stays the same.

IV. EXPERIMENTS

A. DATA COLLECTION

This paper adopts the Indiana University Chest X-ray Collection, which is a subset of Open-i (Open Access Biomedical Image Search Engine). The data set contains 3,955 radiology reports and 7,470 chest X-rays, including 2,314 abnormal reports, accounting for 58.51%, and each includes the indication, Findings, Impression, and Manual annotation section. After counting all the reports, 517 reports were found missing the Findings section, and 34 were found missing the Impression section. The missing parts are marked as "Findings data is null" or "Impression data is null" respectively. The Manual annotation is annotated and cross-validated by many radiologists [22]. In the process of manual annotation, each label corresponds to a radiological discovery. Meanwhile, Demner-Fushman *et al.* [5] used the Medical Subject Headings (MeSH)[2] vocabulary to standardize the processing of Findings and Impression from each radiology report. Among which 6,519 terms were assigned to 2,314 abnormal reports, ranging from 1 to 13 terms per the report, and close to

[2] https://www.nlm.nih.gov/mesh/meshhome.html

3 on average. This work eliminated the ambiguity of different doctors' description of the same symptoms, which used in this paper as the ground truth for automatic annotate.

B. TASKS

The research work of this paper is divided into three subtasks: abnormal annotation, multiple disease annotation and complete Annotation.

Task 1 Abnormal Annotation: For the most basic task, it is only judged whether the generated label is normal or not, and it does not distinguish whether the generated exception is the same as the actual exception. When the model generation is annotated as *"normal"*, it is judged to be normal, whereas all the other results are considered as abnormal. In this task, it can be considered as text classification work. We refer to the work of Dong *et al.* [23], using the results of K-Medoids clustering as the label of the report.

Task 2 Multiple Disease Annotation: Further, annotation the disease contained in the report. We refer to Mostafiz and Ashraf [7] previous work and use the Named Entity Recognition(NER) method for pathology terms extraction. Among the 23 pathology terms selected as disease labels are: *"opacity"*, *"aorta"*, *"fractures"*, *"osteophyte"*, *"scoliosis"*, *"density"*, *"pneumothorax"*, *"cardiomegaly"*, *"emphysema"*, *"arthritis"*, *"granuloma"*, *"kyphosis"*, *"pneumonia"*, *"spondylosis"*, *"deformity"*, *"hypertension"*, *"consolidation"*, *"mass"*, *"thickening"*, *"hernia"*, *"lucency"*, *"consolidation"*, and *"bronchiectasis"*. In the neural model, attention is paid only to whether or not the above 23 pathology terms are included in the generated result, and the words related to the disease description is ignored.

Task 3 Complete Annotation: The complete annotation is considered as a special summary, which represents the relationship between the positions where the term appears. So, we need to determine the similarity between the generated and the reference annotation. The pointer-generator is a well-known abstraction generation neural model and also the baseline we refer to. The details of the pointer-generator model can be found in [15]. More importantly, the impact of the indication information on the generated annotation is evaluated by comparing different fusion methods.

C. METRICS

We use different metrics depending on the task. In the abnormal annotation and multiple disease annotation tasks, the Precision, Recall, and F1 score are used to evaluate the performance. The P, R, and F1 values are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

TABLE 1. Binary classification on normal vs. Abnormal, where P-G represents the pointer-generator model (the same below).

Models	Accuracy	Precision	Recall	F1
K-Medoids	82.70%	82.33%	80.11%	81.20%
P-G	95.87%	94.55%	93.13%	93.83%
P-G $\oplus$ Indication	93.63%	92.90%	92.06%	92.73%
Our model	**96.26%**	**95.12%**	**93.60%**	**94.35%**

In the Mostafiz and Ashraf [7] study, the pathology terms extracted from the report are also considered as a sequence, and the extraction results are evaluated using the BLEU value. BLEU [24] is often used to determine the similarity between two sentences which has four levels of assessment accuracy. The indicators are distinguished by a hierarchy of accuracy from low to high: BLEU-1, BLEU-2, BLEU-3, and BLEU-4. Regardless of the BLEU accuracy, the full score of 1.0 indicates that the two sentences are completely matched, while 0.0 indicates they are irrelevant. We also use BLEU to evaluate Task 2 and Task 3 separately. In addition, we also use another commonly used metric ROUGE-L [25] in natural language. Further, we will explore an evaluation metric that is more suitable for medical report annotations in Section V.

D. DETAILS

All data are divided into a training set, verification set, and test set according to the proportion 80%, 10%, 10% respectively. In the experiment of K-Medoids. we define the similarity of clauses based on the edit distance [26]. The edit distance is defined by the minimum operations (insert, delete, and replace) that convert one clause to another. At the same time, the k-medoids algorithm [27] is used to perform clustering on clauses. K-medoids is related to the k-means [28] algorithm and selects points in the dataset as cluster centers.

In the neural model, we implements the model by using PyTorch framework. For the sake of training the model, word2vec [29] is used to train the word vector and add $< SOS >$ and $< EOS >$ as the start and end identifiers of the input sentence. In terms of model details, the Encoder Bi-LSTM has a hidden size of 100, while Decoder LSTM has a hidden size of 200, and the vector size of Sentence Embedding is 100. Therefore, Adam optimizer [30] was used to optimize the negative log-likelihood loss.

V. RESULTS AND DISCUSSIONS

A. ABNORMAL ANNOTATION RESULT

We evaluated the Baseline Models and our models on the testset. For the two-class subtasks that distinguish between normal and abnormal, we use {0, 1} to normalize the output. We mark the output "normal" as 0, and the rest of the output "abnormal" is marked as 1. The accuracy, Precision, Recall and F1 scores were calculated separately. Tab. 1 shows a comparison of the results.

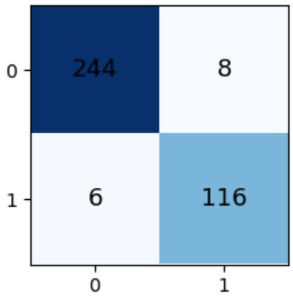


FIGURE 3. Confusion matrix of our model on the testset.

All methods have achieved an accuracy of more than 80%, the lowest of which is K-Medoids. Our model has the best performance and the accuracy rate is 96.26%. As a non-neural model, K-Medoids adopts an unsupervised learning method, which can only rely on the similarity of text to judge, and the precision is not high. The neural model is used as a supervised method. During the training process, the "normal" label can guide the learning of the model, so the effect is much better than K-Medoids. Although the experiment results show that there is little difference between the three neural models, our model still achieves the best results. In addition, Our model's confusion matrix is showed in Fig. 3. Generally, despite different input and model structure, the specially trained neural model is more in line with our automatic annotation goals.

B. MULTIPLE DISEASE ANNOTATION RESULT

We compare the effects of different methods of pathology terms. The NLU[3] is the Natural Language Understanding service provided by IBM, which can analyze and find out the following key concepts from a given text: Concept, Category, Emotion, Entities, Keywords, Relations, Semantic Roles and Sentiments. Tab. 2 shows Precision, Recall, F1 scores calculated for different annotation methods and our model. The results show that the generic tool NLU has the worst effect on entity extraction for domain-specific text. The overall effect of the end-to-end model is higher than the NER method. Our model shows the best results, with precision, recall and f1 scores of 52.87%, 55.04% and 53.93% respectively.

TABLE 2. Precision, recall and F1 scores calculated for multiple disease annotation result.

Models	Precision	Recall	F1 Score
NLU	21.46%	34.55	26.47%
NER [7]	45.34%	55.51	49.91%
P-G	51.11%	53.74%	52.19%
P-G ⊕ Indication	48.54%	49.93%	49.22%
Our model	**52.87%**	**55.04%**	**53.93%**

At the same time, we show the BLUE and ROUGH-L values in Tab. 3. The extracted words for each report were joined to form a sentence. Each sentence was considered as an annotation for a report. Similarly, the extraction of the NLU tools is the worst. Our model works best, which the scores of BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE-L are 65.12, 17.13, 9.69, 3.36 and 56.26, respectively. Regardless of which metric is used, the end-to-end deep neural model in this task is better than the general-purpose model and the NER method. Comparing the fusion of different indication information, directly splicing the indication to the completion of the finding and impression will cause the effect to deteriorate. Our model for the individual encode of indication can better reflect the reasoning relationship between indication and finding.

TABLE 3. Multiple disease annotation result, where B-1, B-2, B-3, B-4, R-L represent BLEU-1, BLEU-2, BLEU-3, BLEU-4 and ROUGE-L, respectively.

Models	B-1	B-2	B-3	B-4	R-L
NLU [7]	45.05	3.96	0.39	0.04	-
NER [7]	49.53	4.82	0.48	0.05	-
P-G	61.88	16.49	8.21	2.43	51.22
P-G ⊕ Indication	59.12	15.02	7.77	1.99	46.16
Our model	**65.12**	**17.13**	**9.69**	**3.36**	**56.26**

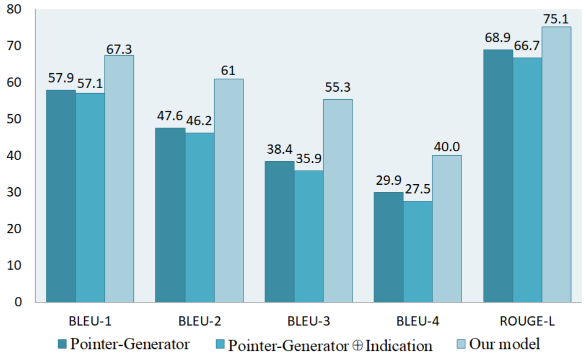


FIGURE 4. BLEU and ROUGE-L scores calculated for complete annotation result on the testset.

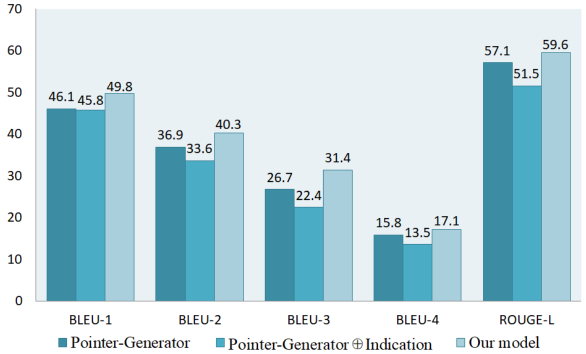


FIGURE 5. BLEU and ROUGE-L scores calculated for complete annotation result on the abnormal testset.

C. COMPLETE ANNOTATION RESULT

The results of abnormal and multiple disease Annotation show that the end-to-end deep neurological model works

[3] https://console.bluemix.net/catalog/services/natural-language-understanding

Finding	Normal heart size mediastinal contours. No focal airspace consolidation. No hyperexpansion of the lungs. No pleural effusion or pneumothorax.	Left midlung opacity noted, not visualized on prior. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures intact.	**Low lung volumes are present.** The heart size and pulmonary vascularity appear within normal limits. There has been interval development of bibasilar opacities. The appearance of the right base opacity XXXX atelectasis. The left base opacities could represent early pneumonia or areas of atelectasis. No pneumothorax or pleural effusion is seen.	There is chronic asymmetric elevation of the right hemidiaphragm. Compared with the prior study, there is mildly increased streaky airspace disease in the right lung base. Hilar prominence appears stable. There is no pneumothorax or large pleural effusion. Heart size is stable and grossly normal. There no acute bony findings.
Impression	No acute cardiopulmonary abnormality.	Left mid lung opacity noted, most compatible with atelectasis versus infiltrate. Recommend clinical correlation.	1. Low lung volumes. 2. XXXX XXXX opacities. Right base appears to represent atelectasis. Left base could be atelectasis or pneumonia.	Chronic asymmetric elevation of the right hemidiaphragm with mildly increased right basilar airspace disease, atelectasis versus infiltrate.
Indications	The patient is a XXXX-year-old female with XXXX and XXXX. History of asthma.	XXXX-year-old male, shortness of breath, question pneumonia.	Low oxygen saturation.	Shortness of breath.

Human	normal	Opacity / lung / lingula;
		Pulmonary Atelectasis / lingula;
		Infiltrate / lung / lingula
			Lung / hypoinflation;
			Opacity / lung / base / bilateral;
			Pulmonary Atelectasis / base / right
				Diaphragm / right / elevated / chronic;
				Airspace Disease / lung / base /right / streaky;
				Pulmonary Atelectasis / base / right

Pointer-Generator	normal	opacity / lung / base / left;
		pneumothorax / mild
			lung / pulmonary;
			Opacity / lung / base / bilateral;
			pneumonia/ base / left
				pneumothorax / right / paratracheal / prominent;
				airspace disease / lung / base / right;

Pointer-Generator ⊕ Indication	normal	opacity / lung / base / left;
		pneumothorax / lingula
			atelectasis / pulmonary ;
			Opacity / lung / base / left;
			pneumonia/ base / left
				breath / right /;
				pneumothorax /effusion / mild;
				hemidiaphragm;

Our Model	normal	opacity / lung / lingula;
		pulmonary atelectasis / base / left;
		Infiltrate / mild
			lung / **hypoinflation**;
			opacity / lung / base /bilateral;
			pneumonia / base / bibasilar ;
			atelectasis
				Hemidiaphragm / right / elevation / chronic;
				airspace disease / lung /base /right ;
				pneumothorax /base/right

FIGURE 6. Comparison of generated annotation results by different models.

better than the non-neural model, so in the annotations containing the disease and disease description, we only evaluate the deep neural model. Fig. 4 shows the comparison between the evaluation results of different models. Our model has distinct advantages. Furthermore, the BLEU-4 and ROUGE-L values are 20.1% and 6.2% higher than the Pointer-Generator model respectively, indicating that we will significantly optimize the effect of automatic annotation by adding the input of indication to the basic Pointer-Generator model.

Similarly, the the Pointer-Generator ⊕ Indication model uses direct splicing indication to cause the model input information to be too long, which ultimately affects the annotation effect of the model. *"XXXX-year-old male with chest pain"* is an example. Obviously, the description of *"chest pain"* can only be used as a basis for judging abnormalities, because it is not a radiological term. Actually, by directly connecting the indication information and findings and impressions, the input of the model is increased, the noise of the model is increased, and the effect of the original model is reduced In contrast, we optimize the model's probability distribution in the decoder part by inputting indications that are processed through sentence embedding in the middle of the model, and the final results show that the method is more reasonable and effective.

The result of Sec. V-A shows that the accuracy rate is over 90% in the judgment of the abnormality. The normal labels are all "normal". For more objective evaluation, we exclude normal samples from the testset and only retain the abnormal samples. Fig. 5 shows the results of the evaluation of the abnormal samples.

The output results of different models are shown in Fig. 6. Human represents manual annotation, and the red font part represents the same effect as a manual annotation. In this way, we can learn that the neural model can accurately determine whether the report is abnormal or not. Meanwhile, we noticed that the neural model can automatically generate words that do not appear in the report as annotations. The bold part of Fig. 6 shows that the description of *"Low lung volumes are present,"* which is consistent with the *"Lung/hypoinflation"* label in the report. However, as a technical term, "hypoinflation" does not appear in the report, which is the standardization problem we mentioned in section 1 that needs to be solved through the neural model.

D. NEW EVALUATION METRICS

In the analysis of the results, we found that the use of the evaluation metrics like BLEU-4 and ROUGE-L cannot be objective enough in the task of automatic generating annotations. For example, for the reference label *"density/lung/base/right/mild,"* two different prediction labels *"density/lung/base/right"* and *"lung/base /right/mild"* have the same BLEU-4 and ROUGE-L values. However, in fact, the former misses a degree word *"mild,"* while the latter misses a key term *"density."* Obviously, the former reflects more of the original report, despite the fact that both of them have one word less than the reference label.

The conventional BLEU-4 and ROUGE-L evaluation metrics do not reflect the primary and secondary relationship between words and words from annotations, which is a special summary of the report. On the basis of traditional evaluation methods, a more proper metric for annotation is proposed by adding the evaluation of the first word. Equation 16 shows our calculation method.

$$M - ROUGE = (\mathbb{1}\{r_1 = p_1\} + ROUGE(R, P))/2 \qquad (16)$$

R is the reference label, $R = \{r_1, r_2, \ldots, r_N\}$, N is the reference label length, $N > 1$. P is the prediction label $P = \{p_1, p_2, \ldots, p_M\}$, M is the prediction label length, $M > 1$. Where $\mathbb{1}$ is the indication function, $ROUGE$ refers to the calculation method of [31].

Calculated with the new evaluation metric, the values of *"density /lung/base/right"* and *"lung/base/right /mild"* are 0.925 and 0.425, respectively. The new metric is called M-ROUGE, and we think this way can better distinguish the primary and secondary relationships in the complete annotation. Of course, the new metric suggests that more validation is needed, and then we will validate this metric in a larger data set and ask professional radiologists for an auxiliary assessment.

VI. CONCLUSION

In this article, we try to use the neural network model to dig out the key information from the natural language of radiology reports and automatically annotate radiological images. we combine the work of entity extraction, relationship extraction, and standardization in traditional annotation tasks. Based on the depth model generated by the general summary, we propose an end-to-end automatic annotation generation model. We tested the effects of our model on a public dataset. The results show that our method has a significant improvement in baseline compared to the three sub-tasks of abnormal, Multiple Disease, and complete Annotation. In addition, we consider the influence of indication on annotations. After comparing experiments with different indication information fusion methods, it is found that the use of indication information can better guide the model decode. Finally, we tentatively proposed a new evaluation index M-ROUGE, which can better evaluate the radiology report labeling effect by improving ROUGE-L.

REFERENCES

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[2] A. R. Aronson, J. Mork, F.-M. Lang, W. Rogers, A. Jimeno-Yepes, and J. C. Sticco, "The NLM indexing initiative: Current status and role in improving access to biomedical information," U.S. Nat. Library Med., Bethesda, MD, USA, Tech. Rep. TR-2012-001, 2012.

[3] G. T. Gobbel *et al.*, "Assisted annotation of medical free text using RapTAT," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 833–841, 2014.

[4] L. Tonin, "Annotating mentions of coronary artery disease in medical reports," KTH Roy. Inst. Technol., Stockholm, Sweden, Tech. Rep., 2017.

[5] D. Demner-Fushman, S. E. Shooshan, L. Rodriguez, S. Antani, and G. R. Thoma, "Annotation of chest radiology reports for indexing and retrieval," in *Multimodal Retrieval in the Medical Domain*. Vienna, Austria: Springer, 2015, pp. 99–111.

[6] X. Zhou, B. Liu, Z. Wu, and Y. Feng, "Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks," *Artif. Intell. Med.*, vol. 41, no. 2, pp. 87–104, 2007.

[7] T. Mostafiz and K. Ashraf, "Pathology extraction from chest X-ray radiology reports: A performance study," 2018, *arXiv:1812.02305*. [Online]. Available: https://arxiv.org/abs/1812.02305

[8] R. Leaman, R. Khare, and Z. Lu, "Challenges in clinical natural language processing for automated disorder normalization," *J. Biomed. Inform.*, vol. 57, pp. 28–37, Oct. 2015.

[9] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, 2010.

[10] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019, *arXiv:1901.07031*. [Online]. Available: https://arxiv.org/abs/1901.07031

[11] I. Banerjee, M. C. Chen, M. P. Lungren, and D. L. Rubin, "Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort," *J. Biomed. Inform.*, vol. 77, pp. 11–20, Jan. 2018.

[12] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," 2015, *arXiv:1509.00685*. [Online]. Available: https://arxiv.org/abs/1509.00685

[13] R. Nallapati *et al.*, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," 2016, *arXiv:1602.06023*. [Online]. Available: https://arxiv.org/abs/1602.06023

[14] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," 2016, *arXiv:1609.07843*. [Online]. Available: https://arxiv.org/abs/1609.07843

[15] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," 2017, *arXiv:1704.04368*. [Online]. Available: https://arxiv.org/abs/1704.04368

[16] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," 2017, *arXiv:1705.04304*. [Online]. Available: https://arxiv.org/abs/1705.04304

[17] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," 2018, *arXiv:1805.11080*. [Online]. Available: https://arxiv.org/abs/1805.11080

[18] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2497–2506.

[19] J. M. Bosmans, J. J. Weyler, A. M. De Schepper, and P. M. Parizel, "The radiology report as seen by radiologists and referring clinicians: Results of the COVER and ROVER surveys," *Radiology*, vol. 259, no. 1, pp. 184–195, 2011.

[20] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*. [Online]. Available: https://arxiv.org/abs/1508.04025

[21] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[22] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2015.

[23] Y. Dong, Y. Pan, J. Zhang, and W. Xu, "Learning to read chest X-ray images from 16000+ examples using CNN," in *Proc. 2nd IEEE/ACM Int. Conf. Connected Health, Appl., Syst. Eng. Technol.*, Jul. 2017, pp. 51–57.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[25] C.-Y. Lin, "Rouge: A package automatic evaluation of summaries," in *Proc. ACL Workshop*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

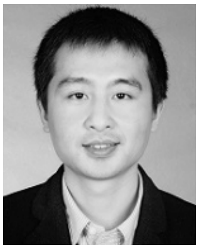
[26] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, 2001.

[27] L. Kaufman and P. J. Rdusseeun, "Clustering by means of medoids," Dept. Math. Inform., Delft Univ. Technol., Delft, The Netherlands, Tech. Rep., 1987, vol. 87003.

[28] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

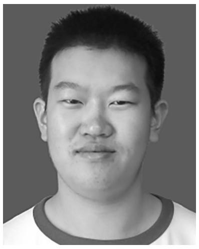
[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980



XIN HUANG was born in 1984. He received the M.S. degree from Nanchang University, in 2010. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, China. His research interests include image processing, data fusion, machine learning, and intelligent systems with applications to medicine.

YU FANG received the Ph.D. degree from Tongji University, China, in 2006, where she is currently a Professor with the Department of Computer Science and Technology. Her current research interests include big data analytics and intelligent systems with applications to medicine.



MINGMING LU was born in 1991. He received the B.S. degree in computer science from the China University of Mining and Technology, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, China. His current research interests include machine learning, natural language processing, and intelligent systems with applications to medicine.

YAO YAO graduated from Tongji University. She received the bachelor's degree in computer science and technology, in 2017. She is currently pursuing the M.S. degree with Tongji University. Her current research interests include natural language processing and text retrieval serving for medical field.

MAOZHEN LI received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, in 1997. He is currently a Professor with the Department of Electronic and Computer Engineering, Brunel University London, U.K. He has more than 160 research publications in these areas, including four books. His current research interests include high performance computing, big data analytics, and intelligent systems with applications to smart grid, smart manufacturing, and smart cities. He is a Fellow of the British Computer Society and the IET. He has served more than 30 IEEE conferences and is on the Editorial Board of a number of journals.