information

MDPI

Article

IKAR: An Interdisciplinary Knowledge-Based Automatic Retrieval Method from Chinese Electronic Medical Record

Yueming Zhao, Liang Hu and Ling Chi *

College of Computer Science and Technology, Jilin University, Changchun 130012, China
* Correspondence: chiling@jlu.edu.cn

Abstract: To date, information retrieval methods in the medical field have mainly focused on English medical reports, but little work has studied Chinese electronic medical reports, especially in the field of obstetrics and gynecology. In this paper, a dataset of 180,000 complete Chinese ultrasound reports in obstetrics and gynecology was established and made publicly available. Based on the ultrasound reports in the dataset, a new information retrieval method (IKAR) is proposed to extract key information from the ultrasound reports and automatically generate the corresponding ultrasound diagnostic results. The model can both extract what is already in the report and analyze what is not in the report by inference. After applying the IKAR method to the dataset, it is proved that the method could achieve 89.38% accuracy, 91.09% recall, and 90.23% F-score. Moreover, the method achieves an F-score of over 90% on 50% of the 10 components of the report. This study provides a quality dataset for the field of electronic medical records and offers a reference for information retrieval methods in the field of obstetrics and gynecology or in other fields.

Keywords: information retrieval; natural language processing; decision support model; ultrasound report; obstetrics and gynecology

Citation: Zhao, Y.; Hu, L.; Chi, L. IKAR: An Interdisciplinary Knowledge-Based Automatic Retrieval Method from Chinese Electronic Medical Record. *Information* **2023**, *14*, 49. https://doi.org/10.3390/info14010049

Academic Editor: Thomas Mandl

Received: 21 November 2022
Revised: 5 January 2023
Accepted: 11 January 2023
Published: 13 January 2023

1. Introduction

Electronic medical records (EMRs) from millions of patients have become significant sources of useful clinical data over the last few decades [1]. With the development of natural language processing technology, information retrieval plays an important role in the healthcare field. It has two main applications, one of which is to improve the efficiency of hospital examination processes. For example, Chen and Émilien et al. [2,3] used deep learning to alleviate the problem of overcrowding in emergency departments. Viincenza et al. [4] used natural language processing techniques and machine learning models to facilitate the digitization of medical prescriptions. Another is to assist doctors in diagnosing diseases. Roch, Sances, and Li et al. have built NLP systems to diagnose pancreatic cysts [5], headaches [6], and pediatric disorders [7]. Although information retrieval is now widely used in the medical field, to the best of our knowledge, few studies have focused on obstetrical and gynecological ultrasound reports.

The ultrasound report, as an essential component of the EMR, serves as the primary means of communication between the sonographer and the doctor who provides the final diagnosis [8]. The use of ultrasound is essential in the clinical practice to examine gynecological diseases and fetal conditions. Many women may choose to have regular ultrasound examinations, which has led to a rapid increase in the number of ultrasound reports. However, ultrasonic diagnoses are typically made by depending on physicians' experience, which might result in limited efficiency. With the rapid increase of massive ultrasound reports, applying specific algorithms to process medical texts becomes necessary. The current studies about the diagnosis of radiology reports are mainly aimed at areas such as liver cancer [9] and breast cancer [10] and concerned with the classification of reports [11] or extracting recommendations from reports [12,13] etc. To the best of our knowledge, there

are relatively few scholars who focus on the automated diagnosis from Chinese ultrasound reports in obstetrics and gynecology.

However, processing Chinese ultrasound reports is more challenging compared to other languages due to the complexity of Chinese. Such challenges are mainly reflected in the following aspects:

- Suitable ultrasound reports in obstetrics and gynecology are difficult to obtain. Training the model requires a large number of labeled datasets. Available Chinese datasets for public access are not found on the internet.
- The majority of ultrasound reports are unstructured text. Despite the fact that a considerable quantity of relevant medical data is saved, the lack of a standard structural framework and the existence of many flaws, such as improper grammar, spelling errors, and semantic ambiguity, make data processing and analysis more difficult.
- Chinese does not use spaces to separate words, which is different from English. Therefore, named entity recognition is a key initial step in natural language processing of Chinese [14,15]. However, named entity recognition based on a general corpus is less effective.
- Traditional diagnostic approaches depend primarily on the healthcare professional's judgment, which might be subjective at times. Two doctors may make different diagnoses based on their expertise and experience if there is no gold standard or predetermined level of agreement on diagnostic criteria [16].
- As the diagnosis is an inferred result based on the doctor's knowledge and experience, words may appear in the ultrasound diagnosis that do not appear in the ultrasound descriptions. Three cases of ultrasound reports are shown in Table 1: (1) diagnostic results can be extracted directly within the report; (2) diagnostic results are not in the report and are obtained by inference; (3) part of the diagnostic results is in the report and the other part is not. The left column shows the ultrasound descriptions summarized by the sonographer from the ultrasound images, and the right column shows the diagnostic result. The texts in red are the words used in ultrasonic descriptions and ultrasound diagnosis that are basically the same. The text in blue is the ultrasound diagnosis as summarized by the corresponding ultrasonic descriptions that does not use the original words in the ultrasonic descriptions.

In the first row of Table 1, the ultrasound diagnosis extracted the sentence "Uterus anteverted with normal size" from the ultrasound description as the conclusion because the patient's findings were normal. In the second row of Table 1, the phrase "hypoechoic nodules" indicates that the patient may have a myoma. In addition, "in the anterior uterus wall" means the location of the myoma, and therefore, the conclusion is "fibroid" based on inference. In the third row of Table 1, "pressure traces can be seen on the fetal neck" indicates that the umbilical cord is wrapped around the fetal neck, and the "U" shape means the umbilical cord is wrapped around the neck once, so the conclusion is inferred to be "fetal cord wrapped around the neck once". At the same time, all other findings of the fetus were normal, so the ultrasound diagnosis extracted the statement "single live fetus" from the ultrasound description as the conclusion.

The ultrasound diagnosis is made on the basis of the ultrasound description, which amounts to a second review of the ultrasound description by the sonographer and is a waste in terms of time efficiency. Thus, this paper targets the diagnostic problems of Chinese ultrasound reports in obstetrics and gynecology. The research objectives of this paper are as follows: (1) to develop an effective deep learning model for large-scale, rapid diagnosis to reduce the workload of sonographers, and (2) to build a publicly available dataset of obstetrical and gynecological ultrasound reports and to find some effective ways to address the above challenges.

Table 1. Examples in the dataset.

Ultrasonic Descriptions	Ultrasound Diagnosis
子宫前位，正常大，宫腔线清，内膜厚1.2 cm，子宫肌层回声均匀。双卵巢正常大。CDFI：未见异常血流信号。 (Uterus anteverted with normal size. The endometrium is clearly visible. No abnormality seen in the myometrial echo. Endometrial thickness is 1.2 cm.The size of left and right ovary is normal. CDFI: No abnormal blood flow signal was observed.)	子宫正常大 (The size of the uterus is normal.)
子宫前位，正常大，宫腔线清，内膜厚0.5 cm，宫壁回声不均匀。子宫前壁见4.0*4.1 cm低回声结节。双卵巢正常大，CDFI：未见异常血流信号。 (Uterus anteverted with normal size. The endometrium is clearly visible. Endometrial thickness is 0.5 cm. Abnormality seen in the myometrial echo. A 4.0*4.1 cm hypoechoic nodule is seen in the anterior uterus wall. The size of left and right ovary is normal. CDFI: No abnormal blood flow signal was observed.)	子宫肌瘤 (Fibroid)
宫内单胎。胎位: 头位; 胎心: 145次/分。AFI: 10 cm; BPD: 5.2 cm; HC: 19.0 cm; AC: 17.2 cm; FL: 3.7 cm。胎盘: 前壁0级。胎儿描述: 头颅:颅骨光环完整,其内结构未见明显异常。眼:可见。上唇: 皮肤回声连续,未见明显异常。四腔心:可见。胃泡: 显示。膀胱: 显示。双肾: 大小正常。胎儿其他情况: 胎儿颈部可见“U”形压迹。孕妇情况: 双卵巢未显示，双附件区未及明显包块。 Single viable fetus. VERTEX presentation at the time of scan. Fetal heart rate is about 145 B/m. AFI: 10 cm; BPD: 5.2 cm; HC: 19.0 cm; AC: 17.2 cm; FL: 3.7 cm.Placenta on anteverted aspect. Grade 0. Fetal description: Head: The cranial halo is intact and its structure shows no obvious abnormalities. Eyes: Visible. Upper lip: continuous skin echogenicity, no significant abnormalities. Four-chamber view of normal fetal heart: Visible. Magenblase: Visible. Urinary bladder: Visible. Kidney: Normal size. Other conditions of the fetus: “U”-shaped pressure traces can be seen on the fetal neck. The situation of pregnant women: The size of left and right ovary is normal. No obvious masses in the bilateral annex area.)	宫内单活胎，胎儿脐带绕颈一周，请结合临床。 (Single viable fetus. Fetal umbilical cord wrapped around the neck once. Please correlate with clinical finding.)

To achieve these purposes, we built a new automated diagnostic model that extracts key information from ultrasound reports and automatically generates diagnostic results. In the first step, we used desensitized data from the ultrasound department of the Second Affiliated Hospital of Jilin University to create a dataset containing complete reports (including ultrasound descriptions and diagnostic results) for 180,000 patients. In the second step, the dataset was preprocessed to remove typos and redundant text to facilitate subsequent use of the dataset by other researchers. In the third step, a specialized lexicon in the field of ultrasound was established to effectively improve the accuracy of named entity recognition. In a fourth step, a sequence-to-sequence model is used, while a pattern-matching algorithm is added to extract the relationship between ultrasound descriptions and diagnostic findings, with the aim of addressing the problem of words that do not appear in ultrasound descriptions appearing in diagnostic findings. Finally, we propose the synonym processing method and probabilistic accuracy methods that effectively reduce the influence of physicians' subjective thinking.

In summary, the contributions of our paper are summarized as follows:

- We constructed and published a fully open dataset containing 180,000 Chinese obstetric and gynecologic ultrasound reports. Our dataset is available in GitHub https://github.com/rachel12308/Chinese-OB-GYN-report-dataset (accessed on 6 June 2022).

- We proposed an interdisciplinary knowledge-based automatic retrieval method (IKAR) for obstetric and gynecological ultrasound in which the ultrasound diagnosis can be generated automatically from ultrasonic descriptions. The model was applied on the hospital dataset for the experimental verification of its effectiveness and efficiency. As a result, it was proved that the model could achieve an accuracy, recall and F-score of around 90%.
- We have carried out a detailed analysis of the dataset and proposed several targeted approaches to address the challenges encountered in the Chinese diagnostic task. Both of these methods are better at reducing errors and significantly improving inference performance.

The remainder of this paper is organized as follows: Section 2 serves as an introduction of a relevant model or system for information retrieval in the medical field. In Section 3, a public ultrasound report dataset in Chinese is established. In Section 4, the IKAR method is proposed. The details of each module are explained in the related subsections of Section 4. In Section 5, the proposed IKAR method is verified experimentally through the dataset. The performance of the proposed model is then evaluated and compared to some traditional models. Section 6 contains the conclusion of this paper.

2. Related Work

Information retrieval techniques have been widely used in the medical field for the detection of tumors, circulatory system diseases, digestive system diseases and neurological diseases [1]. The main methods utilized are traditional NLP models (rule-based algorithms, self-designed algorithms, etc.), traditional machine learning and neural network models. Except for neural network models, all of the above methods are interpretable and widely used. Neural network models have also gradually come to the attention of researchers in recent years due to their excellent information retrieval capabilities. Table 2 shows the information retrieval methods applied to selected disease types.

Table 2. Disease areas and implementation methods for medical information retrieval.

Application Areas	Methods	No. of Papers
Circulatory system diseases	cTAKES + Self-designed NLP Algorithm [17] Rule-based algorithm [18] KnowledgeMap Concept Identifier (KMCI) [19] MedTagger + Self-designed NLP Algorithm [20] CUIMANDREef + Self-designed Algorithm [21] CNN VS traditional NLP models [22]	7
Nervous system diseases	RF/SVM/LR/CNN + Rule-based Algorithm [23] MetaMap+Build Dictionary [24]	2
Digestive system diseases	Unstructured Information Management Architecture (UIMA) + Rule-based Algorithm [5] Self-designed NLP Algorithm [25]	2
Tumor	Neural Network model [26] 13 Neural Network models [27]	1
Other	Transformer-based model [28] BERT [29] CheXbert [30]	4

Fu et al. [23] built a system that utilized both rule-based and machine learning approaches. This system is used to identify Silent Brain Infarction (SBI) and White Matter Disease (WMD) from electronic health records (EHR), and the accuracy rate can exceed 0.9. Selen et al. [31] present a model for natural language processing that combines a rule-based feature extraction module with a conditional random field model. The model can extract 96% accurate measurements and core characteristics from radiology reports. Zhou et al. [24] utilized an NLP approach to extract lifestyle information from clinical record data for

260 sick and healthy persons. They explored the factors that might cause AD dementia based on this knowledge. According to the findings, the approach accurately extracts 74% of the influencing factors. Warner et al. [32] used an NLP algorithm to extract cancer stage information from electronic health records. The result showed that 72% of patients could identify the specific stage (e.g., stage I, stage II). Mehrabi et al. [33] proposed a rule-based NLP method to identify patients with a family history of pancreatic cancer. On two public datasets, the method achieves 87.8% and 88.1% accuracy, respectively. Farrugia et al. [34] developed an NLP system for extracting cancer stage and recurrence information from radiological reports. This approach has a 97.3% accuracy in identifying original tumor flow, metastasis, and recurrence.

With the development of deep learning, many researchers have started to explore the application of neural networks to medical datasets. Matthew et al. [22] compared the performance of convolutional neural networks (CNN) and traditional NLP models in extracting pulmonary embolism (PE) results. Kenneth et al. [26] also used a neural network model to extract cancer information and identified specific information more accurately. More recently, some more complex neural network models, such as Transformer [35] and BERT [36], have achieved very good results on many NLP tasks. Ignat et al. [27] compared 13 supervised classifiers, including Transformer, and showed that the BiLSTM approach based on the attention mechanism performed the best. David et al. [28] proposed a Transformer-based model for MRI neuroradiology reports with classification performance higher or slightly lower than that of domain experts. Keno et al. [29] used BERT to identify the most important findings in critical care chest radiograph reports and achieved the best performance in identifying congestion, effusion, consolidation, and pneumothora compared with previous methods. Akshay et al. [30] proposed CheXbert to extract one or more clinical findings from radiology reports. The model goes beyond the previous best rule-based tagger and sets up a new SOTA on MIMIC-CXR [37].

Many recent works have demonstrated the usability and effectiveness of neural network models. However, most of the studies are for medical reports in English, perhaps because most of the English reports have standard Unified Medical Language System (UMLS), so there are more corresponding tools available and shared. In the Chinese language domain, there is a lack of standard data formats, so algorithms need to be designed to meet the needs in conjunction with the specific form of the dataset.

3. Dataset Construction

3.1. Data Collection

The dataset was collected from the Second Affiliated Hospital of Jilin University in Changchun, China, between 2012 and 2015, and contains 180,000 ultrasound reports. All identifying information has been removed to protect patient privacy. All reports were approved by the local institutional review board, and informed consent was obtained. All ultrasound reports are unstructured and written in Chinese.

3.2. Preprocessing

In order to effectively analyze the text and reduce the impact of typos and redundant text, the ultrasound report must be preprocessed. The details are described below.

- Step 1. Dealing with typos in texts.
- Step 2. Dealing with redundant texts in the report. Ultrasound is only an ancillary item to help the doctor make a diagnosis. As a result, there are many suggestive phrases in the ultrasound report, such as “re-evaluation after 2–3 weeks” or “please correlate with clinical finding”. These statements were not useful to the clinical support system in our task, so 56 similar suggestive statements were eliminated.

3.3. Named Entity Recognition

Many Asian languages, such as Chinese and Japanese, do not use spaces to separate words, unlike English. Therefore, for a better analysis of the Chinese report, Named Entity

Recognition is a key initial step. In order to improve the accuracy of NER, a specialized dictionary in the field of in obstetric and gynecological ultrasound needs to be created for the following two reasons:

- First, ultrasound reports contain a large number of medical terms. Because the word frequency of specialized vocabulary is much lower than that of common vocabulary, the NER may make mistakes. For example, it is possible to split the sentence "宫腔线清 (The endometrium is clearly visible)" into "宫腔" and "线清", but the correct result is "宫腔线" and "清".
- Second, ultrasound reports are relatively similar in content and often use repeated words. Only 3763 words are used in the ultrasound descriptions of this dataset, while only 498 words are used in the ultrasound findings.

By analyzing the texts of the report, we found that 93.2% of the professional phrases were made up of two or three words, and all phrases of four words or more were made up of short words. The percentages of words of different lengths are shown in Table 3. Therefore, we combine unsupervised and supervised learning methods for NER. Figure 1 shows the process of NER. For unsupervised learning, first, all the words are combined according to the bigram and trigram algorithm and sorted by the number of occurrences. For supervised learning, ultrasonographers were invited to label 2000 representative ultrasound reports. These reports were fed into the BiLSTM-CRF model for training, and the number of occurrences of all words was counted. BiLSTM-CRF [38] is a classical NER model that can significantly improve the performance of Chinese NER tasks, reaching State-of-the-Art on many datasets [39,40]. The model is relatively simple to implement and fast to train. Moreover, the NER task in this paper is only used as an aid to enrich the words in the dictionary, so the BiLSTM-CRF model is chosen. Finally, 581 phrases were selected for the dictionary, combining suggestions from sonographers and terminology from 《Ultrasonographic Diagnosis in Obsterics and Gynecology》 [41].

Table 3. Percentage of three types of words in the dictionary.

Type	Percentage
Words formed by two Chinese characters	76.70%
Words formed by three Chinese characters	16.50%
Words formed by more than three Chinese characters	6.80%

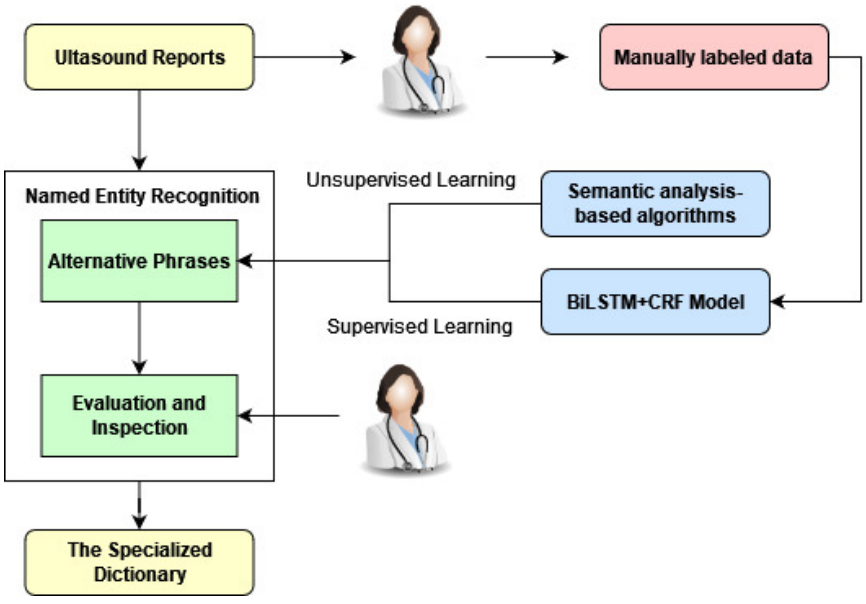

Figure 1. The process of building a dictionary.

To transform sentences into a form that the computer can understand, ultrasonic descriptions are segmented via a Chinese tokenizer with the dictionary. The accuracy rate of NER increased from 88% to 95%. Moreover, the dictionary covers 94.2% of all Chinese characters in the ultrasound reports.

3.4. Data Analysis

A total of 29,816 ultrasound reports were analyzed in this paper, in which 16 reports are blank and therefore excluded. In this section, 19,900 reports are put in the training set, 4950 reports are put in the validation set and 4950 reports are put in the test set. Table 4 shows the word count statistics of the dataset after segmentation and correction. "Diagnosis" denotes the number of reports. "Tokens_description" and "Tokens_result" indicate the total number of words after tokenization of the ultrasound description and ultrasound result, respectively. Each ultrasound description has an average of 39 words and each ultrasound result has an average of four words.

Table 4. Word count statistics of the training and test sets.

	#Number	#Tokens_DESCRIPTION	#Tokens_RESULT
Train	19,900	767,026	81,304
Dev	4950	195,843	20,342
Test	4950	196,976	20,949
Total	29,800	1,159,845	122,595

4. IKAR Method

4.1. Task

The task of this paper is to generate specific phrases from ultrasound descriptions and use them as diagnostic results. The model takes in sentences from ultrasound descriptions as inputs and outputs critical phrases. Figure 2 illustrates the task of this paper, which aims to automatically extract specific information to help the doctor make a diagnosis.

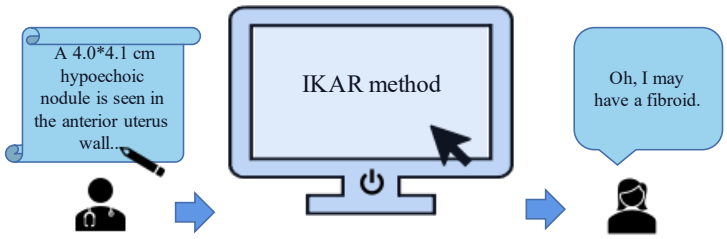

Figure 2. The task of the paper.

4.2. Implementation of the System

The IKAR method takes advantage of both the Transformer model and relation extraction method. After analyzing the diagnosis report and consulting the experts in obstetrics and gynecology, we developed the Synonyms Handling method and Probabilistic Accuracy algorithm to further enhance the performance of the system. The overall view of the system is shown in Figure 3. Each process of the model is explained in detail in the following subsections.

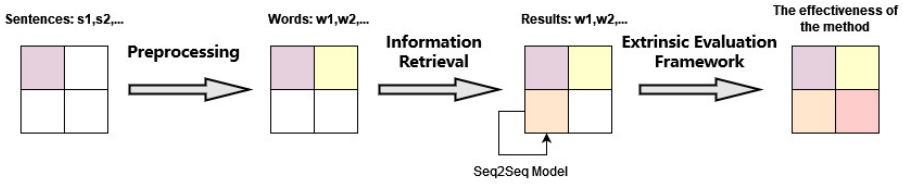

Figure 3. Overall view of the system.

4.2.1. Word Embedding

In order to transform words into a form that the computer can understand, words need to be transformed into word vectors. Glove, Word2Vec and BERT are common word-embedding models. In this work, the BERT model is chosen for word embedding. BERT is context dependent, and the same word in different contexts may generate different representations. Moreover, BERT takes into account the position of the word in the sentence.

4.2.2. Sequence-to-Sequence Model

Sequence-to-sequence models have been widely used in the fields of machine translation and keyword generation, and they have achieved good results. In the field of machine translation, the seq2seq models deal with text in two languages, such as translating a sentence from English to Chinese. In the field of keyword generation, the input data to the model is usually a piece of news or a paper, and the output is the keywords of the article. The task in this paper has some commonalities with the tasks described above. The three cases shown in Table 1 may also exist in the above task. As a result, the sequence-to-sequence (seq2seq) model was chosen for information retrieval as a solution to the problem of ultrasound results containing terms that do not appear in the ultrasound descriptions. Figure 4 represents the data flow of this part.

In the first step, the ultrasound descriptions and ultrasound findings from the training set are fed into the transformer model for training. In the second step, the ultrasound descriptions from the test set are fed into the model trained in the first step to obtain preliminary results. In the third step, some error results are modified according to the relation extraction algorithm to obtain the final results of this part.

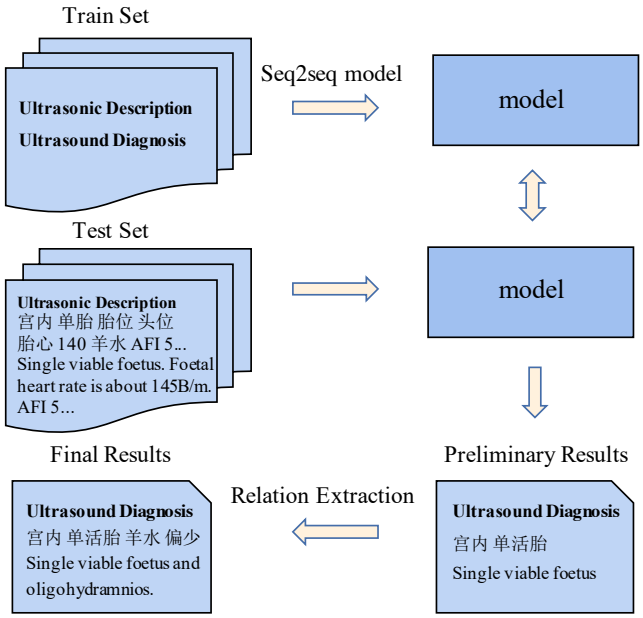

Figure 4. Pipeline of the information retrieval.

Transformer was selected as the model underlying the information retrieval task. It is a fully attention-based encoder–decoder model that uses a multi-headed attention approach. Transformer considers a different focus for each word when generating the resulting sequence, substantially improving the accuracy of the prediction. In the realm of machine translation, the advent of the attention mechanism marks a watershed moment. Bahdanau et al. [42] proposed an attention-based model in 2014. Instead of a fixed-length vector, the encoder in the model maps the source text as a sequence of vectors. At each time step, the decoder selects a subset of these vectors to generate tokens in the target sentence. Google proposed the Transformer model based on the idea of multi-headed

attention. Since then, Transformer has gradually become the mainstream model in the field of natural language processing. The improved model based on Transformer has reached state-of-art on several datasets such as WMT2014 English-German [43], IWSLT2015 English-Vietnamese [44], DUC 2004 Task 1 [45], etc. The attention mechanism fits the characteristics of the task in this paper. Each sentence in the ultrasound result corresponds to a different part of the ultrasound description. Therefore, the Transformer model is chosen for knowledge extraction in this paper.

The Transformer model consists of two parts: the encoder and the decoder. The encoder continuously corrects the vector to better represent the semantic information of the context. The decoder uses the attention mechanism when generating words, reasoning about the words that should be output later based on the previous output. The decoder continuously performs the inference process until it inferred the end marker <EOS>, which indicates the end of the task. The training is stopped after five consecutive rounds. At this point, the ultrasound descriptions of the test set are input to the model, and the initial prediction results of the model are obtained.

4.2.3. Relation Extraction

By analyzing the initial prediction results obtained from the Transformer model, the pattern-matching algorithm was chosen for relation extraction in this paper. The pattern-matching algorithm has the advantages of high accuracy, customized for specific domains, easy to implement and simple to build [46]. Therefore, three patterns are defined based on the experimental results as shown in Table 5. If the sentences satisfying the following patterns can be extracted from the ultrasound description, the corresponding diagnostic results can be obtained directly. For example, the "胎儿颈部可见"U"形压迹 ("U"-shaped pressure marks can be seen on the fetal neck)" in the ultrasound description fits the pattern in the first row of Table 5—"胎儿 (fetal)" + "颈部 (neck)" + "可见 (can see)" + ""U"形压迹" ("U"-shaped pressure marks). Thus, the diagnostic result is "胎儿脐带绕+颈+一周 (fetal umbilical cord wrapped around the + neck + one times)".

Table 5. Definition of patterns.

Definition of Patterns	Result
胎儿+位置+动词+压迹形状 (Fetus + position + verb + shape of pressure marks)	胎儿脐带绕+位置+一周/两周/三周 (Fetal umbilical cord wrapped around the + position + one/two/three times)
AFI + number (number range from 0 to 7.9)	羊水 偏少 (Oligohydramnios)
AFI + number (number greater than 18)	羊水 偏多 (Hydramnion)

The results of the Transformer model are modified according to Table 5. As shown in Algorithm 1, firstly, the statements in the ultrasound descriptions of the test set that fit the patterns in Table 5 were filtered out. Then, we determine whether the above texts have already output the correct result. If the correct results have been output, no modification is made; otherwise, the incorrect results are removed and the correct diagnostic results are added.

4.3. Extrinsic Evaluation Framework

Since there is no standard format for Chinese ultrasound reports, different doctors may have different words to describe the same disease. Therefore, the diagnostic habits of doctors can also influence the accuracy of the results. To mitigate the impact of this problem, we developed and integrated the Synonyms Handling method and Probabilistic Accuracy algorithm in the IKAR method.

4.3.1. Synonyms Handling

Different doctors may use different words to express the same meaning when making a diagnosis. For example, "极少" and "过少" (they both mean little), "偏多" and "过多' (they both mean much), etc. If a correct word in the test set is not inferred, but its synonym is inferred, then the synonym should be considered as a correct prediction. The Word2Vec tool is used to obtain the word vectors after segmenting the ultrasound reports into words. The cosine similarity between "极少" and "过少" was calculated to be 0.953, and the cosine similarity between "偏多" and "过多" was calculated to be 0.968. Therefore, the problem of synonyms in the results of the model and the test set can be handled by the cosine similarity. When the cosine similarity between the model-generated word and the correct word is greater than or equal to 0.9, the model-generated word is considered to be correct.

Algorithm 1: Pattern-Matching Algorithm.

Input : The ultrasound descriptions $(w_1, ..., w_i, ..., w_n)$ in the test set, $i \in [1, n]$
The ultrasound findings $(u_1, ..., u_i, ..., u_n)$ in the test set, $i \in [1, n]$
The result $(p_1, ..., p_i, ..., p_n)$ obtained by inputting $(w_1, ..., w_i, ..., w_n)$ into the seq2seq model, $i \in [1, n]$
Output: The modified result
foreach *w_i of the number i* **do**
  if *keywords in w_i* **then**
    if $u_i = p_i$ **then**
      continue
    else
      delete error words
      modify the p_i according to the rules in Table 5
    end
  end
end

4.3.2. Probabilistic Accuracy

There is no standard format for Chinese ultrasound reports, so different doctors may choose to use different phrases to describe diagnostic results that are normal or not obviously serious. For example, some physicians will give a diagnosis of "子宫正常大 (the size of the uterus is normal.)" for "子宫前位，正常大，宫腔线清，内膜厚 0.8 cm，宫壁回声不均匀。双卵巢正常大。CDFI: 未见异常血流信号。 (Uterus anteverted with normal size. The endometrium is clearly visible. No abnormality seen in the myometrial echo. Endometrial thickness is 0.8 cm. Abnormality seen in the myometrial echo. The size of left and right ovary is normal. CDFI: No abnormal blood flow signal was observed.)", while others may give a diagnosis of "回声不均 (Uneven echogenicity)". From a professional point of view, all of this patient's indicators are normal. The uneven echogenicity of the uterine wall is also normal and does not require additional attention. A standard diagnosis does not exist for this patient's ultrasonic description. At this time, the result of the model is "子宫正常大", "回声不均" or neither of them should be considered as correct prediction.

Algorithm 2 was proposed to alleviate this problem. In the first step, if "回声 不均匀 (Uneven echogenicity)", "回声 不均 (Uneven echogenicity)" or "子宫 正常 大 (The size of the uterus is normal)" appears in both the ultrasound description and ultrasound result of the test set, but the above words do not appear in the output of the model, then the correct words "回声 不均匀", "回声 不均", or "子宫 正常 大" should be added to the model output. In the second step, if "回声 不均匀", "回声 不均" or "子宫 正常 大" appears in both the ultrasound description of the test set and the model output, but the above words do not appear in the ultrasound result of the test set, then "回声 不均匀", "回声 不均" or "子宫 正常 大" is added to the diagnostic results of the test set. This method minimizes the influence of the doctor's diagnostic habits and allows our model to calculate accuracy, recall and F-score more precisely.

4.3.3. Evaluation Metrics

The evaluation method uses accuracy, recall, and F-score. The evaluation objects are the correct diagnostic results in the test set and the prediction results of the model, as shown in Equations (1)–(3).

$$acc = p_{true}/(p_{true} + p_{false}) \tag{1}$$

$$rec = p_{true}/(p_{true} + n_{false}) \tag{2}$$

$$f1 = 2 \times acc \times rec/(acc + rec) \tag{3}$$

In Equations (1) and (2), p_{true} indicates how many words in the correct diagnostic result were correctly predicted by the model. p_{false} indicates how many words in the correct ultrasound result were not predicted by the model. n_{false} indicates how many words in the model's predicted result did not appear in the correct diagnostic result.

Algorithm 2: Probabilistic accuracy.

Input : The ultrasound descriptions $(w_1, ..., w_i, ..., w_n)$ in the test set, $i \in [1, n]$
The ultrasound findings $(u_1, ..., u_i, ..., u_n)$ in the test set, $i \in [1, n]$
The final output $(p_1, ..., p_i, ..., p_n)$ of the module in 3.3.1, $i \in [1, n]$
Output: Modified results $(p_1, ..., p_i, ..., p_n)$ of the model output, $i \in [1, n]$
Modified ultrasound results $(u_1, ..., u_i, ..., u_n)$ of the test set, $i \in [1, n]$

```
foreach w_i of the number i do
    if keywords in w_i and u_i then
        add keywords to p_i
    end
end
foreach w_i of the number i do
    if keywords in w_i and p_i then
        add keywords to u_i
    end
end
```

5. Experiments

In this section, the proposed system's performance was assessed. This section was divided into two parts: (a) results for the sequence-to-sequence models and (b) results for the IKAR method.

5.1. The Basic Sequence-to-Sequence Models

In this study, four representative seq2seq models were considered for comparison with the Transformer model.

- seq2seq+RNN is the earliest seq2seq model [47]. This model sets one RNN as the encoder and one RNN as the decoder, which can score some sequences. This model can also generate target sequences based on source sequences.
- seq2seq+LSTM [48] replaces the RNN module of the above model with an LSTM module. Complete sentences are used for training instead of just phrases.
- seq2seq+copyRnn [49] first applied the encoder–decoder structure to the keyword generation problem. By adding a replication mechanism to the RNN, it helps the model to predict those words with a low number of occurrences in the original text.
- seq2seq+Reinforcement Learning [50] introduces reinforcement learning to the keyword generation task for the first time. The model is set up with an adaptive reward function. The function first uses the recall value as a reward to ensure that a sufficient number of keywords is generated.

The above neural network models mainly used the deep learning framework of PyTorch [51] and Jieba [52]. All models are trained using NVIDIA GeForce RTX 3060. The parameters of the models are set as follows: the batch size is set to 50, the learning rate is set to 1×10^{-4} , the training epoch is set to 10, and the number of steps is 4990. For text greater than 512 in length, the first 512 of the text is retained. The majority of ultrasound reports are relatively short and rarely have more than 512 words. For the occasional text that is too long, the extra-long parts are simply deleted.

Table 6 shows the results of five classical models. The data entered into these five models are words that have been processed only by tokenization without the use of other processing methods involved in IKAR. It can be seen that the LSTM model performs the worst, with nearly 25% less accuracy than the other models, and even worse than the RNN model that appeared first. This indicates that this task does not need to summarize the information of too many phrases in the source text, and there is no long-term dependency problem. The correct result can be inferred by analyzing the words in the vicinity of the keywords. For the copyRnn method with the copy mechanism, the accuracy is not as good, which is probably because many words appear in the ultrasound results that are not in the ultrasound descriptions. The result of the reinforcement learning model is better, but it is not as good as the transformer model, which is completely based on attention. The Transformer model can achieve a neutral sum of accuracy and recall with the highest F-score.

Table 6. Accuracy, recall and F-score of the five basic model.

	Accuracy	**Recall**	**F-Score**
LSTM	58.14%	86.82%	69.64%
RNN	86.20%	88.89%	87.52%
copyRnn	84.97%	91.10%	87.93%
Reinforcement Learning	86.21%	89.63%	88.39%
Transformer	**87.42%**	**90.75%**	**89.05%**

The result may indicate that the Transformer's multi-headed attention mechanism is better suited to the task and can capture important information from the entire sentence. In the ultrasound description shown in Figure 5, the uterus, the endometrium, the intrauterine, the uterine wall, the right annex area, the left ovary and the left annex area are described. In the ultrasound result shown in Figure 5, the abnormalities of the intrauterine and right adnexal regions in the previous text are summarized. "宫内暗区 (Dark area seen in the uterus)" is derived from "宫内见暗区0.9*0.6 cm, 0.8*0.7 cm (Dark areas of 0.9*0.6 cm and 0.8*0.7 cm are seen in the uterine cavity)". "右附件区囊性包块 (Cystic mass visible in the right adnexal region)" is derived from "右附件区见4.2*3.3 cm无回声，界限尚清，形态尚规则 (There is a 4.2*3.3 cm anechoic zone in the right adnexal area. The right adnexal area is well defined and regular in shape)". Therefore, it is necessary to focus on some of the keywords when generating ultrasound diagnostic results. It is consistent with the advantages of the attention mechanism.

5.2. Pattern-Matching Algorithm

The results of the basic sequence-to-sequence model after adding the pattern-matching algorithm are shown in Table 7. It illustrates that the algorithm greatly improves the performance of the model. All models have about a 1% decrease in recall rate and 1% to 7% increase in accuracy rate, thus reaching a higher F-score. It shows that the pattern-matching method proposed in this paper is effective in increasing the number of correct words. The best-performing Transformer model eventually achieves an accuracy, recall and F-score value of about 90%.

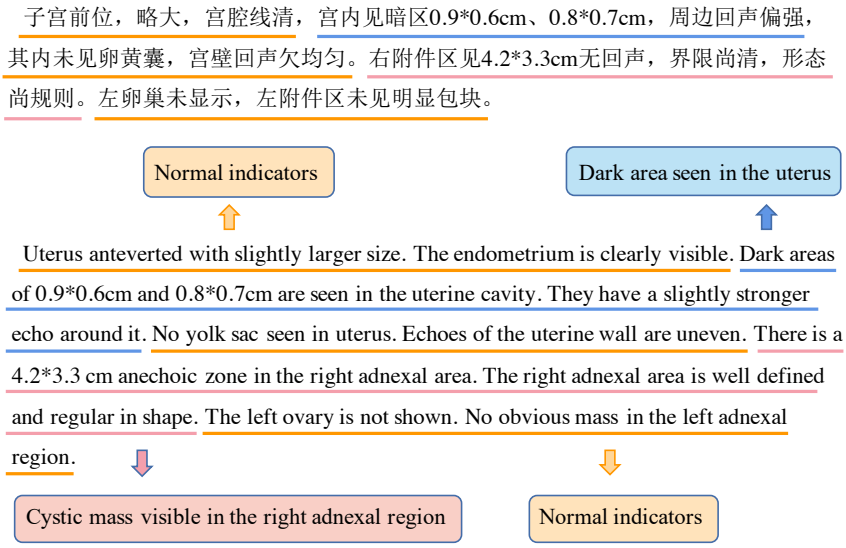

Figure 5. Example of a pair ultrasound description and finding.

Table 7. Accuracy, recall and F-score of the five basic model with pattern-matching algorithm.

	Accuracy	Recall	F-Score
LSTM	63.00%	85.71%	72.62%
RNN	87.03%	88.04%	87.53%
copyRnn	87.32%	90.10%	88.69%
Reinforcement Learning	88.23%	89.22%	88.72%
Transformer	**89.38%**	**91.09%**	**90.23%**

As can be seen from Table 7, the IKAR approach, which is the Transformer model combined with the pattern-matching algorithm, performs best in the evaluation framework proposed in this paper. The method proposed achieves an accuracy of 89.38%, recall of 91.09% and F-score value of 90.23%. It shows that the relation extraction method proposed in this paper is effective in increasing the number of correct words. By processing synonyms and incorporating probabilistic accuracy methods, the method also corrected some phrases that were mistakenly thought to be incorrect.

5.3. Individual Examination Items

To further analyze the performance of the IKAR method, the ultrasound report was subdivided into 10 sections. Table 8 shows the 10 common components of the ultrasound report. These 10 parts cover the entire content of the ultrasound report and are key reference information for doctors to determine whether the patient has gynecological diseases or to check whether the fetal condition is normal.

Figure 6 shows an example of a diagnostic result. According to the classification criteria in Table 8, this sentence can be divided into three parts: the condition of the uterus, the condition of the annex area, and the condition of the ovaries. According to this criteria, we divided the results of our model and the correct results of the test set into 10 parts. Then, the accuracy, recall and F-score were calculated separately for each part. Table 9 shows the test results for each component. Figure 7 plotted the detailed performance of these 10 components based on IKAR method proposed in this paper.

Table 8. Ten common items in ultrasound reports.

	Name	Brief Description
1	The condition of the uterus	Describe the size, presence, shape and number of the uterus, etc.
2	The condition of the annex area	Describe the presence or absence of abnormal masses in the annex areas.
3	The condition of the ovaries	Describe the presence or absence of polycystic ovary syndrome.
4	The condition of the cervix	Describe the presence of lesions or abnormalities in the cervix.
5	The condition of the vagina	Describe the presence of lesions or abnormalities in the vagina.
6	Whether the patient is pregnant	Describe whether the patient is pregnant, confirm fetal viability and check for the number of fetuses.
7	The condition of the fetus	Describe the basic condition of the fetal head, eyes, heart, etc.
8	Fibroid	Describe whether the patient has fibroid and their number.
9	Adenomyosis	Describe whether the patient has adenomyosis.
10	Other abnormal conditions	Describe other abnormalities in the patient's report.

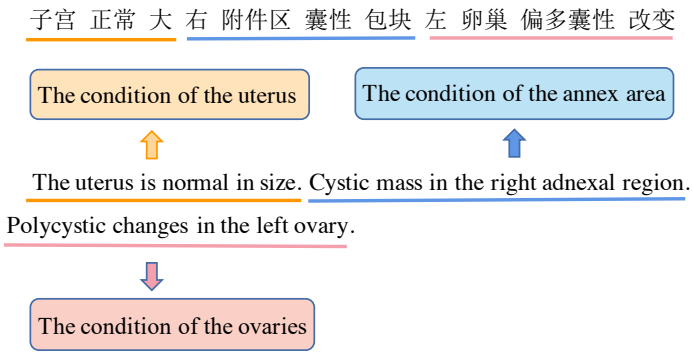

Figure 6. Example of an ultrasound finding.

In Table 9, the F-scores for the condition of the uterus, the condition of the annex area, and the condition of the fetus were all improved by 1% to 12% after processing in combination with the method proposed in this paper. The recall is greater than the accuracy for the condition of the ovaries, the condition of the cervix, fibroid and other abnormal conditions. This indicates that the current method accurately captures certain potential information. For the condition of the uterus, the condition of the vagina, and the condition of the fetus, the accuracy was greater than the recall. This indicates that the current method needs to add more rules to produce correct answers. In Table 9, the current method performs better for six items numbered 1, 4, 5, 6, 8 and 9. This is because the descriptions of these six items are relatively simple, and most of them have an ultrasound result of about three words. The system works well for such items with simple descriptions. On the contrary, the performance of the remaining four items was poor, which may because the ultrasound results of these terms were more complex and involved more words. The accuracy of the fetal condition was significantly lower than the other items, which may be because some disease such as "Tetralogy of Fallot", "Cystic Hygromas" and "Dandy–Walker malformation" were present in the test set but do not appear in the training set or appear less than 10 times. The descriptions of these diseases were all relatively complex, and the model was not able to learn the relevant features well enough to predict these diseases. These erroneous outputs significantly affect the accuracy.

Table 9. Accuracy, recall and F-score of our model for each part.

	Name of Each Part	IKAR		
		Accuracy	**Recall**	**F-Score**
1	The condition of the uterus	**95.56%**	**93.99%**	**94.77%**
2	The condition of the annex area	**79.37%**	**81.00%**	**80.18%**
3	The condition of the ovaries	76.23%	87.72%	81.97%
4	The condition of the cervi	83.17%	95.45%	88.89%
5	The condition of the vagina	99.48%	98.46%	98.96%
6	Whether the patient is pregnant	95.81%	96.88%	96.34%
7	The condition of the fetus	**80.54%**	**62.16%**	**70.17%**
8	Hysteromyoma	86.87%	93.8%	90.2%
9	Adenomyosis	90.48%	90.19%	90.34%
10	Other abnormal conditions	79.19%	90.55%	84.49%

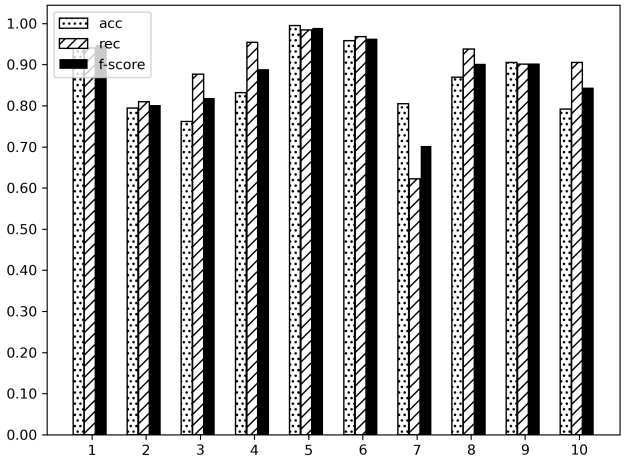

Figure 7. Accuracy, recall and F-score of IKAR for each part.

5.4. Limitations

Although the method proposed in this paper can achieve some effectiveness on the dataset, there are still some limitations. First, since the method proposed in this paper is a baseline experiment for this dataset, only a few classical seq2seq models have been tested. In future work, some other neural network models (e.g., BERT, etc.) can be considered. Second, the pattern-matching algorithm proposed in this paper does not take into account all the rules available in ultrasound reports. This is because the types of ultrasound reports are complicated, and it is difficult to identify the common features of certain diseases. In future work, further consultation with ultrasonographers and continued study of the reports can be performed to extract more association rules. Third, the lack of interpretability of neural network models is also a critical issue in gaining physicians' trust in clinical practice. How to solve this problem is a common direction for scholars' future research.

The publicly available dataset provided in this paper also has some limitations. First, the dataset has a data imbalance problem. The method cannot identify some rare gynecological and obstetric diseases because of the small number of ultrasound reports for some diseases. The inclusion of algorithms to deal with data imbalance can be considered in future work. Moreover, the model may not output the results expected by doctors in other hospitals when applied directly to their datasets due to the different idioms of different doctors. In future work, the diagnosis can be converted to labels by manually labeling the dataset, which in turn converts the information extraction problem into a multi-label classification problem with more general applicability.

6. Conclusions

In this paper, a publicly available dataset of obstetrical and gynecological ultrasound reports is built. This dataset contains 180,000 complete Chinese ultrasound reports. We have processed the errors contained in this dataset and provided a dictionary in the field of ultrasound, which will help interested scholars use this dataset for more meaningful research. Based on this dataset, a new knowledge extraction method (IKAR) is proposed. The IKAR method could generate diagnostic results from Chinese ultrasound reports automatically, so it could serve as a quick and convenient tool to provide a reference in final clinical diagnosis in the domain of gynecology and obstetrics.

In this model, the Transformer model and relation extraction algorithm are used and further developed according to the professional knowledge of gynecology and obstetrics to enhance the performance of data information retrieval in the reports. In addition, the Synonyms Handling method and Probabilistic Accuracy algorithm are developed to reduce the influence of subjective expressions and build a standardized format for ultrasound reports, which further improves the performance of the model.

After a detailed explanation of the model, the proposed IKAR method and four traditional sequence-to-sequence models are applied to a hospital dataset for experimental verification. As a result, the IKAR method has the best performance for obstetric and gynecological ultrasound reports. The IKAR method could achieve 89.38% accuracy, 91.09% recall, and 90.23% F-score. Compared to traditional models such as RNN and LSTM, the IKAR method has an overall improvement of 1% to 8% in accuracy and 1% to 3% in F-score. Among the 10 components of the ultrasound report, 90% of the components achieved an F-score of 80% or more, and 50% of the components achieved 90% or more.

In the future, the characteristics of ultrasound descriptions and the rules of identifying diseases can be further studied in order to improve the accuracy of identifying rare gynecological diseases. By analyzing the reports of other diseases, the IKAR method could have a larger dictionary, and it is possible to diagnose more kinds of diseases and be applied in other medical fields. Finally, the IKAR method is not limited to information retrieval of texts; it also has potential application for the "information retrieval" of images, such as automated medical image recognition to identify the location and properties of lesions from images [53,54]. If integrated with an automated image recognition method, the IKAR method will have potential to perform an entirely automated ultrasound examination.

Author Contributions: Y.Z.: Conceptualization, Methodology, Formal analysis, Investigation, Writing—original draft, Writing—reviewing and editing, Validation. L.H.: Resources, Project administrating. L.C.: Conceptualization, Methodology, Software, Supervision, Resources, Writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the National Key R&D Plan of China (grant number 2017YFA0604500), by the National Sci-Tech Support Plan of China (grant number 2014BAH02F00), by the National Natural Science Foundation of China (grant number 61701190), by the Youth Science Foundation of Jilin Province of China (grant number 20160520011JH, 20180520021JH), by the Youth SciTech Innovation Leader and Team Project of Jilin Province of China (grant number 20170519017JH), by the Key Technology Innovation Cooperation Project of Government and University for the whole Industry Demonstration, China (grant number SXGJSF2017-4), by the Key scientific and technological R&D Plan of Jilin Province of China (grant number 20180201103GX), and by the Project of Jilin Province Development and Reform Commission, China (grant number 2019FGWTZC001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our dataset is available in GitHub: https://github.com/rachel12308/Chinese-OB-GYN-report-dataset (accessed on 6 June 2022).

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

Abbreviations

The following abbreviations are used in this manuscript:

RF	Random Forest
EMR	Electronic Medical Record
SVM	Support Vector Machine
LR	Logistic Regression
CNN	Convolutional Neural Networks
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System
NER	Named Entity Recognition

References

1. Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S.; et al. Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **2018**, *77*, 34–49. [CrossRef] [PubMed]
2. Chen, C.H.; Hsieh, J.G.; Cheng, S.L.; Lin, Y.L.; Lin, P.H.; Jeng, J.H. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. *Int. J. Med. Inform.* **2020**, *139*, 104146. [CrossRef] [PubMed]
3. Arnaud, É.; Elbattah, M.; Gignon, M.; Dequen, G. Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4836–4841.
4. Carchiolo, V.; Longheu, A.; Reitano, G.; Zagarella, L. Medical prescription classification: A NLP-based approach. In Proceedings of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), Leipzig, Germany, 1–4 September 2019; pp. 605–609.
5. Roch, A.M.; Mehrabi, S.; Krishnan, A.; Schmidt, H.E.; Kesterson, J.; Beesley, C.; Dexter, P.R.; Palakal, M.; Schmidt, C.M. Automated pancreatic cyst screening using natural language processing: A new tool in the early detection of pancreatic cancer. *HPB* **2015**, *17*, 447–453. [CrossRef] [PubMed]
6. Sances, G.; Larizza, C.; Gabetta, M.; Bucalo, M.; Guaschino, E.; Milani, G.; Cereda, C.; Bellazzi, R. Application of bioinformatics in headache: The I2B2-pavia project. *J. Headache Pain* **2010**, *11*, S134–S135.
7. Li, X.; Wang, H.; He, H.; Du, J.; Chen, J.; Wu, J. Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks. *BMC Bioinform.* **2019**, *20*, 62. [CrossRef]
8. Cai, T.; Giannopoulos, A.A.; Yu, S.; Kelil, T.; Ripley, B.; Kumamaru, K.K.; Rybicki, F.J.; Mitsouras, D. Natural language processing technologies in radiology research and clinical applications. *Radiographics* **2016**, *36*, 176–191. [CrossRef]
9. Liu, H.; Xu, Y.; Zhang, Z.; Wang, N.; Huang, Y.; Hu, Y.; Yang, Z.; Jiang, R.; Chen, H. A natural language processing pipeline of chinese free-text radiology reports for liver cancer diagnosis. *IEEE Access* **2020**, *8*, 159110–159119. [CrossRef]
10. Castro, S.M.; Tseytlin, E.; Medvedeva, O.; Mitchell, K.; Visweswaran, S.; Bekhuis, T.; Jacobson, R.S. Automated annotation and classification of BI-RADS assessment from radiology reports. *J. Biomed. Inform.* **2017**, *69*, 177–187. [CrossRef]
11. Lakhani, P.; Kim, W.; Langlotz, C.P. Automated detection of critical results in radiology reports. *J. Digit. Imaging* **2012**, *25*, 30–36. [CrossRef]
12. Yetisgen-Yildiz, M.; Gunn, M.L.; Xia, F.; Payne, T.H. A text processing pipeline to extract recommendations from radiology reports. *J. Biomed. Inform.* **2013**, *46*, 354–362. [CrossRef]
13. Dutta, S.; Long, W.J.; Brown, D.F.; Reisner, A.T. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. *Ann. Emerg. Med.* **2013**, *62*, 162–169. [CrossRef] [PubMed]
14. Peng, F.; Feng, F.; McCallum, A. Chinese segmentation and new word detection using conditional random fields. In Proceedings of the COLING 2004: 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004; pp. 562–568.
15. Zheng, X.; Chen, H.; Xu, T. Deep learning for Chinese word segmentation and POS tagging. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 647–657.
16. Schiff, G.D.; Hasan, O.; Kim, S.; Abrams, R.; Cosby, K.; Lambert, B.L.; Elstein, A.S.; Hasler, S.; Kabongo, M.L.; Krosnjar, N.; et al. Diagnostic error in medicine: Analysis of 583 physician-reported errors. *Arch. Intern. Med.* **2009**, *169*, 1881–1887. [CrossRef] [PubMed]
17. Savova, G.K.; Fan, J.; Ye, Z.; Murphy, S.P.; Zheng, J.; Chute, C.G.; Kullo, I.J. Discovering peripheral arterial disease cases from radiology notes using natural language processing. In Proceedings of the AMIA Annual Symposium Proceedings. American Medical Informatics Association, Washington, DC, USA, 13–17 November 2010; Volume 2010, p. 722.
18. Tian, Z.; Sun, S.; Eguale, T.; Rochefort, C.M. Automated extraction of VTE events from narrative radiology reports in electronic health records: A validation study. *Med. Care* **2017**, *55*, e73. [CrossRef] [PubMed]
19. Hinz, E.R.M.; Bastarache, L.; Denny, J.C. A natural language processing algorithm to define a venous thromboembolism phenotype. In Proceedings of the AMIA Annual Symposium Proceedings. American Medical Informatics Association, Washington, DC, USA, 16–20 November 2013; Volume 2013, p. 975.

20. Afzal, N.; Sohn, S.; Abram, S.; Liu, H.; Kullo, I.J.; Arruda-Olson, A.M. Identifying peripheral arterial disease cases using natural language processing of clinical notes. In Proceedings of the 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, USA, 24–27 February 2016; pp. 126–131.
21. Kim, Y.; Garvin, J.; Heavirland, J.; Meystre, S.M. Improving heart failure information extraction by domain adaptation. In *MEDINFO 2013*; IOS Press: Amsterdam, The Netherlands, 2013; pp. 185–189.
22. Chen, M.C.; Ball, R.L.; Yang, L.; Moradzadeh, N.; Chapman, B.E.; Larson, D.B.; Langlotz, C.P.; Amrhein, T.J.; Lungren, M.P. Deep learning to classify radiology free-text reports. *Radiology* **2018**, *286*, 845–852. [CrossRef]
23. Fu, S.; Leung, L.Y.; Wang, Y.; Raulli, A.O.; Kallmes, D.F.; Kinsman, K.A.; Nelson, K.B.; Clark, M.S.; Luetmer, P.H.; Kingsbury, P.R.; et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Med. Inform.* **2019**, *7*, e12109. [CrossRef]
24. Zhou, X.; Wang, Y.; Sohn, S.; Therneau, T.M.; Liu, H.; Knopman, D.S. Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. *Int. J. Med. Inform.* **2019**, *130*, 103943. [CrossRef]
25. Ludvigsson, J.F.; Pathak, J.; Murphy, S.; Durski, M.; Kirsch, P.S.; Chute, C.G.; Ryu, E.; Murray, J.A. Use of computerized algorithm to identify individuals in need of testing for celiac disease. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e306–e310. [CrossRef]
26. Kehl, K.L.; Elmarakeby, H.; Nishino, M.; Van Allen, E.M.; Lepisto, E.M.; Hassett, M.J.; Johnson, B.E.; Schrag, D. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol.* **2019**, *5*, 1421–1429. [CrossRef]
27. Drozdov, I.; Forbes, D.; Szubert, B.; Hall, M.; Carlin, C.; Lowe, D.J. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. *PLoS ONE* **2020**, *15*, e0229963. [CrossRef]
28. Wood, D.A.; Lynch, J.; Kafiabadi, S.; Guilhem, E.; Al Busaidi, A.; Montvila, A.; Varsavsky, T.; Siddiqui, J.; Gadapa, N.; Townend, M.; et al. Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). In Proceedings of the Medical Imaging with Deep Learning, PMLR, Lima, Peru, 4 October 2020; pp. 811–826.
29. Bressem, K.K.; Adams, L.C.; Gaudin, R.A.; Tröltzsch, D.; Hamm, B.; Makowski, M.R.; Schüle, C.Y.; Vahldiek, J.L.; Niehues, S.M. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics* **2020**, *36*, 5255–5261. [CrossRef] [PubMed]
30. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv* **2020**, arXiv:2004.09167.
31. Bozkurt, S.; Alkim, E.; Banerjee, I.; Rubin, D.L. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *J. Digit. Imaging* **2019**, *32*, 544–553. [CrossRef]
32. Warner, J.L.; Levy, M.A.; Neuss, M.N.; Warner, J.L.; Levy, M.A.; Neuss, M.N. ReCAP: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. *J. Oncol. Pract.* **2016**, *12*, 157–158. [CrossRef]
33. Mehrabi, S.; Krishnan, A.; Roch, A.M.; Schmidt, H.; Li, D.; Kesterson, J.; Beesley, C.; Dexter, P.; Schmidt, M.; Palakal, M.; et al. Identification of patients with family history of pancreatic cancer-Investigation of an NLP System Portability. *Stud. Health Technol. Inform.* **2015**, *216*, 604. [PubMed]
34. Farrugia, H.; Marr, G.; Giles, G. Implementing a natural langugage processing solution to capture cancer stage and recurrence. In Proceedings of the European Congress of Radiology-RANZCR-AOCR 2012, Sydney, Australia, 30 August–2 September 2012.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
37. Johnson, A.E.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
38. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
39. Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Liu, S. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 182–192.
40. Dai, X.; Karimi, S.; Hachey, B.; Paris, C. Using similarity measures to select pretraining data for NER. *arXiv* **2019**, arXiv:1904.00585.
41. Xie, H. *Ultrasonographic Diagnosis in Obsterics and Gynecology*; People's Medical Publishing House: Beijing, China, 2005.
42. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
43. Takase, S.; Kiyono, S. Lessons on parameter sharing across layers in transformers. *arXiv* **2021**, arXiv:2104.06022.
44. Vaage, A.B.; Tingvold, L.; Hauff, E.; Van Ta, T.; Wentzel-Larsen, T.; Clench-Aas, J.; Thomsen, P.H. Better mental health in children of Vietnamese refugees compared with their Norwegian peers-a matter of cultural difference? *Child Adolesc. Psychiatry Ment. Health* **2009**, *3*, 1–9. [CrossRef] [PubMed]
45. Takase, S.; Kiyono, S. Rethinking perturbations in encoder-decoders for fast training. *arXiv* **2021**, arXiv:2104.01853.
46. Wang, Y.; Mehrabi, S.; Sohn, S.; Atkinson, E.J.; Amin, S.; Liu, H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 73. [CrossRef] [PubMed]
47. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.

48. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*. 2014. Available online: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf (accessed on 15 February 2020).
49. Meng, R.; Zhao, S.; Han, S.; He, D.; Brusilovsky, P.; Chi, Y. Deep keyphrase generation. *arXiv* **2017**, arXiv:1704.06879.
50. Chan, H.P.; Chen, W.; Wang, L.; King, I. Neural keyphrase generation via reinforcement learning with adaptive rewards. *arXiv* **2019**, arXiv:1906.04106.
51. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. 2019. Available online: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf (accessed on 15 February 2020).
52. Sun, J. Jieba (Chinese for "to Stutter") Chinese Text Segmentation: Built to Be the Best Python Chinese Word Segmentation Module. Available online: https://github.com/fxsjy/jieba (accessed on 15 February 2020).
53. Arimura, H. *Image-Based Computer-Assisted Radiation Therapy*; Springer: Berlin/Heidelberg, Germany, 2017.
54. Gupta, K.K.; Dhanda, N.; Kumar, U. A comparative study of medical image segmentation techniques for brain tumor detection. In Proceedings of the 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 14–15 December 2018; pp. 1–4.