Medical Entity Relation Verification with Large-scale Machine Reading Comprehension

Yuan Xia[1†], Chunyu Wang[1†], Zhenhui Shi[1], Jingbo Zhou[1*], Chao Lu[1], Haifeng Huang[1], Hui Xiong[2]
[1]Baidu Inc., China. [2]Rutgers University, USA.
[1]{xiayuan,wangchunyu03,shizhenhui,zhoujingbo,luchao,huanghaifeng}@baidu.com, [2]hxiong@rutgers.edu

ABSTRACT

Medical entity relation verification is a crucial step to build a practical and enterprise medical knowledge graph (MKG) because high-precision medical entity relation is a key requirement for many MKG-based applications. Existing relation verification approaches for general knowledge graphs are not designed for considering medical domain knowledge, although it is central to achieve high-quality entity relation verification for MKG. To this end, in this paper, we introduce a system for medical entity relation verification with large-scale machine reading comprehension. The proposed system is tailored to overcome the unique challenges of medical relation verification including high variants of medical terms, the high difficulty of evidence searching in complex medical documents, and the lack of evidence labels for supervision. To deal with the problem of variants of medical terms, we introduce a synonym-aware retrieve model to retrieve the potential evidence implicitly verifying the given claim. To better utilize the medical domain knowledge, a relation-aware evidence detector and a medical ontology-enhanced aggregator are developed to improve the performance of the relation verification module. Moreover, to overcome the challenge of providing high-quality evidence due to the lack of labels, we introduce an interactive collaborative-training method to iteratively improve the evidence accuracy. Finally, we conduct extensive experiments to demonstrate that the performance of our proposed system is superior to all comparable models. We also demonstrate that our system can significantly reduce the annotation time by medical experts in real-world verification tasks. It can help to improve the efficiency by nearly 300%. In particular, our system has been embedded into the Baidu Clinical Decision Support System.

CCS CONCEPTS

• **Information systems** → *Information extraction*; *Chemical and biochemical retrieval*; • **Applied computing** → *Health informatics*.

KEYWORDS

Fact Verification; Relation Extraction; Clinical Decision Support

[†]Equal contribution. [*]Corresponding author.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467144

ACM Reference Format:
Yuan Xia[1†], Chunyu Wang[1†], Zhenhui Shi[1], Jingbo Zhou[1*], Chao Lu[1], Haifeng Huang[1], Hui Xiong[2] . 2021. Medical Entity Relation Verification with Large-scale Machine Reading Comprehension. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467144

1 INTRODUCTION

With the rapid development of data-driven medical knowledge graph (MKG) construction, an evidence-based and effective medical relation verification framework is demanding. Information extraction (IE) has notably facilitated the construction of MKG which makes it possible to extract a large number of medical entity relations from medical documents such as online medical websites and Electronic Medical Records (EMRs). However, most of the relations extracted from such medical documents are unverified since the data-driven IE method usually cannot provide the evidence to support the claim (i.e. the relation between two entities). Such unverified relations are always untrustworthy and even unacceptable in medical domain applications. In real-world medical applications, such as Baidu Clinical Decision Support System (CDSS) [1], the interpretable and evidence-based result is necessary for assisting the doctor to make a diagnosis [27]. The verification system should provide concrete evidence to verify the relation is correct, not just the probability. In industrial applications, companies often hire a large number of medical domain experts to annotate and verify every extracted medical entity relations from the knowledge graph. This process is labor-intensive, time-consuming and expensive. Therefore, how to automatically verify the extracted medical relations as well as to improve the efficiency of the verification process becomes a vital problem for building a practical and enterprise MKG. Recently, many research efforts have been devoted to Fact Verification (FV) [9, 11, 14, 16, 24, 25, 31, 32], which aims to verify given claims with the evidence retrieved from plain text. However, most of the existing methods are general frameworks for fact verification, without considering the special properties of medical documents and handling the unique challenges in medical domain, they are less effective when dealing with medical entity relation verification.

There are mainly three challenges when dealing with medical fact verification in industrial applications. The first challenge is that there is a lot of variants for medical terms, especially for medical synonyms. For instance, medical term *abdominal pain* can also refer to as *stomachache, collywobbles, etc.* It becomes more difficult to handle such synonym variations for fact verification in medical documents. The sentence retrieval from medical documents needs to be aware of that the candidate sentence implicitly contains the

[1]https://01.baidu.com/

target entity or the synonym of the target entity. As shown in Table 1, let's say when we are checking the medical claim "stomachache is the symptom of gastritis ", which can be represented with the knowledge graph (KG) triplet form (*gastritis*, *stomachache*, *symptom*), and assuming there is a sentence in the clinical textbook which said "Acute gastritis can have pain or an uncomfortable feeling in their upper abdomen". The sentence selection module in general fact verification systems fails to retrieve this sentence as evidence because they are unable to realize the synonym of the target entity is implicitly contained in this sentence.

The second challenge is how to better utilize medical domain knowledge for medical entity relation verification. When verifying the medical entity relation, the experts with medical domain knowledge can quickly focus on the right part of the paragraphs or sentences and can infer the medical relation with medical ontology knowledge. For instance, when checking the medical relation between *pneumonia* and *chest radiograph*, the experts will focus on the paragraph under the title of *laboratory inspection*, while checking the relation between *pneumonia* and *cough*, the experts will locate at the paragraph under the title of *clinical manifestation*. If the *cough* is the symptom of *lobar pneumonia*, the experts can infer the *cough* is also the symptom of *pneumonia*. However, to the best of our knowledge, the general verification models are unable to perceive the above domain knowledge in the medical domain.

The third challenge is how to provide high-quality evidence due to the lack of evidence labels for supervision. In the medical domain, it is difficult to improve the accuracy of the retrieved evidence since we do not have a huge amount of the labeled data which indicates the evidence is correct or not. In the general fact verification framework, the label of the retrieved evidence can be annotated by crowdsourcing, such as Amazon Mechanical Turk. While it becomes very difficult due to the high domain knowledge requirement for medical domain.

Aiming to improve the accuracy and efficiency of medical relation checking, we introduce a framework for automatic medical entity relation verification with large-scale machine reading comprehension. The overall architecture of the system is illustrated in Figure 1. We develop a three-stage pipeline system: (1) Document Retrieval, a module to narrow our search place and focus on the relevant clinical documents, (2) Synonym-Aware Sentence Selection, a module to select the evidence from the retrieved clinical document, and (3) Machine Reading Comprehension-based Semantic Relation Verification (MSRV), a module to check the medical relations via machine reading comprehension of clinical materials. Additionally, we introduce an interactive collaborative-training method to iteratively improve the accuracy of retrieved evidence.

At first, to handle the first challenge of high variants of medical terms, we propose a synonym-aware sentence selection module that is aware of synonyms of the target entity and can perceive the implicit relations between the target entity and sentences. We first construct the candidate synonym pairs with rule-based and synthetic-based methods, then we construct a synonym prediction sub-task to fine-tune a pre-trained language model which is the ERNIE model [23] in this paper. As ERNIE is a continual pre-training framework for language understanding, after this synonym prediction sub-task, our ERNIE-based sentence selection module can have the ability to tackle the synonym problem.

Claim	(*gastritis*, *stomachache*, *symptom*) Stomachache is the symptom of gastritis.
Evidence	Acute gastritis can have pain or an uncomfortable feeling in their upper abdomen.

Table 1: Example of a Claim-Evidence Pair.

To tackle the second challenge, we first utilize the fine-tuned ERNIE model as an encoder to get the embedding representations of the claim, the evidence, and the evidence metadata. Then, different from previous work like [32], which computes the attention by the claim and the evidence alone, we add a relation-aware matrix to let the evidence detector to sense the correlation between the target entity relation and the evidence metadata. It can realize the global structure of the document and better focus on the context information. After that, we adopt a medical ontology-enhanced aggregator for relation verification.

To overcome the third challenge, we introduce an Interactive Collaborative-Training (ICT) method to iteratively improve the accuracy of retrieved evidence. At each iteration, we train an evidence discriminator to score the retrieved evidence and make predictions on unlabeled data set. Then, we assign the most informative evidence to medical experts for annotation and assign high confidence evidence with refined labels. In the next iteration, we retrain the evidence discriminator with labels annotated by medical experts and train our verification system with unlabeled and labeled evidence dataset. We iteratively process the above procedure, until the accuracy of the evidence meets the requirement.

We summarize our contributions as follows:

- We propose a framework for the automatic verification of the medical entity relation with large-scale machine reading comprehension, which incorporates the authoritative clinical materials for the relation verification. To the best of our knowledge, it is the first deployed system to apply automatic medical relation verification techniques into real-world applications.
- We develop two novel techniques to tackle the challenges for relation verification system in the medical domain: the synonym-aware sentence selection, which captures the synonym and implicit relation in the sentence; the MSRV model, which can make better use of the domain knowledge through a relation-aware evidence detector and a medical ontology-enhanced aggregator.
- We propose an interactive collaborative-training method to tackle the problem of lack of evidence labels in medical relation verification and iteratively improve the accuracy of retrieved evidence, which is significant in real-world medical applications.
- We evaluate our system on offline and online experiments to demonstrate the superiority of our framework with higher performance compared to other comparative models, and our deployed system can also significantly reduce the time of annotation by medical experts in real-world applications.

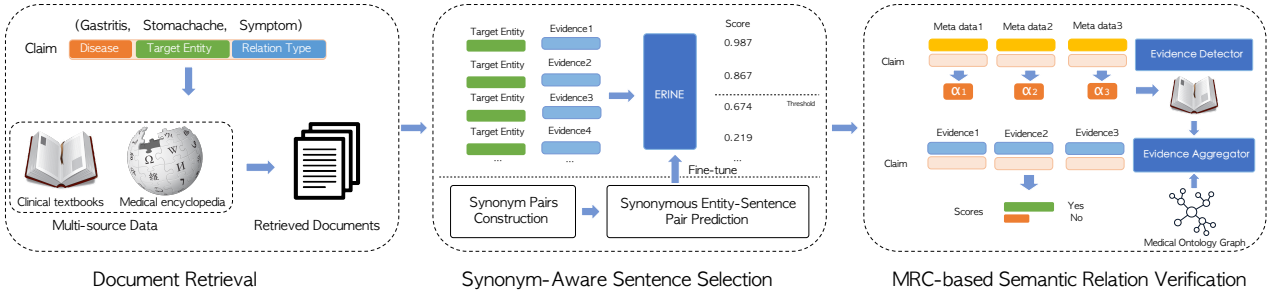

Figure 1: Illustration of our proposed framework for medical entity relation automatic verification.

2 RELATED WORK

Fact verification (FV) is a challenging task that requires retrieving relevant evidence from plain text and utilize the evidence to verify given claims. Existing methods usually formulate FV as a natural language inference (NLI) [1] task. Thorne et al. [24] mainly feed the concatenated evidence and the given claim into the NLI model. Another solution is to utilize the decomposable attention model (DAM) [18] to predict each claim evidence pair individually and then all predictions are aggregated for final verification [15]. There are also a few of studies [9, 12, 29] that adopt the enhanced sequential inference model (ESIM) [5] to infer the relationship between evidence and claims, which achieves better performance. Zhou et al. [32] propose the graph-based evidence aggregating and reasoning (GEAR) model for evidence claim prediction. Liu et al. [14] introduce kernel graph attention network (KGAT) model for fine-grained fact verification with kernel-based attentions. However, the literature mentioned above are general frameworks for fact verification, which are less effective in the medical domain verification. MedTruth [6] proposes a truth discovery method for medical knowledge condition discovery from the multi-source, but this method is mainly to consider relation discovery rather than relation automatic verification.

Pre-trained language models like ERNIE [23], BERT[7], XLNet [28] and OpenAI GPT [19] can achieve huge gains on many Natural Language Processing (NLP) tasks, such as GLUE [26] benchmark, by pre-training on unlabeled corpus and fine-tuning on labeled ones. ERNIE employs transformer encoder and pre-training tasks to fuse bidirectional context information. In our experiments, we utilize the fine-tuned ERNIE model for several subtasks in order to automatically verify medical entity relationship.

In general machine reading comprehension, a benchmark dataset is the Stanford Question Answering Dataset (SQuAD) consisting of over 10k questions posed by crowd workers on Wikipedia articles [20]. Chen et al. [3] propose to tackle open-domain question answering using Wikipedia as the unique knowledge source. Reading and understanding text in the clinical medicine domain is also being a major research problem in the field of NLP. For example, a reading comprehension model SeaReader for question-answering task on clinical medicine is proposed [30]. Chen et al. [4] devise a knowledge abstract matching method to retrieve relevant evidence from medical knowledge base to support medical question answering.

There is also another study to utilize a contextual self-attention multi-scale sentence embedding (CAMSE) model and two scoring strategies to exploit semantic similarity and association between a given question and the corresponding evidence document [10]. Fei et al. [8] propose a hierarchical multi-task word embedding model to learn more representative medical entity embeddings and apply them to medical synonym prediction. Moreover, the authors of [10, 30] are mainly focus on the medical multiple choices question-answering, Fei et al. [8] mainly focus on synonym prediction, which is different from the medical entity relation verification. Niu et al. [17] present an iterative method to generate soft evidence labels for improving the performance of machine reading comprehension, while the accuracy of the generated evidence cannot be guaranteed.

3 METHOD

We propose a three-stage pipeline system to automatically verify the medical entity relations via large-scale machine reading comprehension. Additionally, we introduce an interactive collaborative-training (ICT) method to improve the evidence accuracy. We first describe the process of constructing the document retrieval module with medical structured data. Second, we introduce our synonym-aware sentence selection module, which selects the evidence from the retrieved clinical documents. Finally, we introduce our MRC-based semantic relation verification model (MSRV), which checks the medical entity relations via comprehension of authoritative clinical materials. The full pipeline of our proposed framework for medical entity relation verification is illustrated in Figure 1.

3.1 Document Retrieval

Given a claim, we first retrieve a set of evidence that is likely to be relevant using a document retrieval module. The document retrieval is built upon ElasticSearch, which is a search engine based on the Lucence library followed by BM25 [21]. We parse dozens of clinical textbooks and medical encyclopedia data into 4-dimensional structure (*Disease, Title, Path, Paragraph*) and it serves as the data source for our module. In this paper, we define *Title* and *Path* as *Evidence Metadata*. The details of the clinical materials are described in section 4.1.1. We process all clinical materials into the same data structure as shown in the Table 2. We use entities in the given claim as search queries to find the topK relevant documents. The retrieved documents are then feed into our synonym-aware sentence selection module.

Item	Content
Disease	Cardiac disease
Title	Clinical manifestation
Path	[*Book*]Cardiology → [*Chapter*]Heart Failure → [*Section*]Clinical Manifestations of Heart Failure
Paragraph	Paroxysmal dyspnea often occurs at night, and patients often wake up suddenly during deep sleep, with extreme anxiety and choking ⋯

Table 2: Data Structure for Document Retrieval.

3.2 Synonym-Aware Sentence Selection

In the domain of medical relation verification, the sentence retrieval becomes harder when there are a lot of variations for the medical terms, especially for the case of medical synonyms. As the medical entity normalization is not perfect, the coverage of the available evidence retrieved by ElasticSearch is limited. To solve this problem, we propose a synonym-aware sentence selection module. As the computational complexity for the ERNIE model is expensive, before the ERNIE prediction, we construct a simple semantic similarity model to calculate the relevance score between the candidate sentence and target entities to filter out the irrelevant sentence. In this semantic similarity model, we split the candidate sentence into words, and use the average of word embedding to represent the sentence embedding. The word embedding is trained on medical EMRs with FastText [2, 13]. Then we calculate the similarity of the target entity and sentence using the cosine distance. We filter out the sentences with similarity scores which are below the threshold τ_w, the remained sentences are the candidate set for our synonym-aware ERNIE selection model.

The overall workflow of the synonym-aware sentence selection mainly has following two steps: *synonym pairs construction* and *synonymous entity-sentence pair prediction*.

3.2.1 Synonym Pairs Construction. Due to the limited coverage of the available synonym pairs, we construct additional medical synonym pairs from the existing corpus. Inspired by [8], we use two methods for the synonym pairs construction: a rule-based method for synonym pairs extraction and a synthetic-based method for synonym pairs generation.

For the rule-based method, we define several rule-based templates to extract synonym pairs. For instance, the templates can be *abbreviation for, referred to as, commonly known as, short for, etc.* Then we traverse the corpus with the templates to generate synonym pairs. Additionally, we also parse the medical term encyclopedia in the property field such as *alias attribute*.

For the synthetic-based method, we extract medical synonym pairs based on the existing symptom vocabulary list $\mathcal{V}_s$ and a corresponding attribute vocabulary list $\mathcal{V}_a$. $\mathcal{V}_a$ gives a more specific description of the symptom (such as frequency, intensity, color, duration, location, etc.). The candidate synonym pairs can be constructed by splicing or combining with the symptom and its attributes. For example, by combining the symptom *cough* and attribute word *persistent*, we can get a medical synonym pair (*cough*, *persistent cough*). Note that, although there is a small portion of the synthetic synonym pairs which are not conformed to the actual grammatical rules, the overall influence on the final verification model is insignificant.

Method	Entity	Candidate Sentence
Rule-based	Dental fluorosis	Dental fluorosis is also called *mottling of teeth*.
Synthetic	Cephalalgia	*Paroxysmal headache* for one month, the left frontotemporal area is heavy, ⋯

Table 3: Example of Synonymous Entity-Sentence Set. The text in *italic* is the synonym of the target entity.

3.2.2 Synonymous Entity-Sentence Pair Prediction. In this section, we construct a synonym prediction task to fine-tune the ERNIE model. The objective of the task is to predict whether the target entity or the synonym of the target entity is contained in the given sentence. The detail of corpus construction is described in the section 4.1.2. An example of synonymous entity-sentence pair is shown in table 3.

First, we feed target entity and sentence that contains the synonym of the target into the ERNIE model to obtain the embedding representation.

$$\boldsymbol{T}_s = \text{ERNIE}(e_t, s) \tag{1}$$

where e_t represents the target entity and s represents the candidate sentence contains the synonym entity, $\boldsymbol{T}_s$ is the output representation of the ERNIE model.

Then, we feed the ERNIE embedding representation to a dense layer and a softmax layer to get the final synonym-aware prediction.

$$\text{Score}_\text{s}^{(\text{i})} = \frac{\exp(\boldsymbol{W}_s^\text{T}\boldsymbol{T}_s^{(i)})}{\sum_{j=1}^{C}\exp(\boldsymbol{W}_s^\text{T}\boldsymbol{T}_s^{(j)})} \tag{2}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{C\times F}$ is the output weight matrix, F is the number of hidden dimensions of ERNIE, C is the number of prediction labels (here, $C = 2$), and Score_s is the normalize output probability using the softmax function. Similarly, we set a threshold τ_s to control the trade-off between quality and quantity of the selected evidence with the predicted probabilistic value.

Finally, the evidence retrieved by sentence selection module is fed into our MRC-based semantic relation verification model.

3.3 MRC-based Semantic Relation Verification

Given a medical relation triplet (*Disease Entity*, *Target Entity*, *Relation Type*), and N pieces of retrieved evidence $(e_1, e_2, \cdots, e_N)$, the goal of our machine reading comprehension based semantic relation verification model (MSRV) model is to verify the given

claim with retrieved evidence. In this paper, the medical relation $\mathcal{R}$ can be one of five types: *Symptom*, *Operation*, *Radiographic Examination*, *Laboratory Examination*, and *Others*. Our MSRV model is comprised of two parts: a *Relation-Aware Evidence Detector* for precisely locating the right part of the evidence, and a *Medical Ontology-Enhanced Evidence Aggregator* for final textual entailment based on the medical ontology graph and retrieved evidence.

3.3.1 Relation-Aware Evidence Detector. When verifying the medical entity relation, the experts with medical domain knowledge can quickly focus on the right part of the paragraphs or sentences. However, the general evidence aggregator and verification model are unable to perceive the above medical domain knowledge. We propose a relation-aware evidence detector to focus on the right part of the paragraphs when looking for the evidence.

First, we generate the claim c by concatenating the disease entity e_d, target entity e_t and relation type r ($r \in \mathcal{R}$).

$$c = \text{concat}(e_d, e_t, r) \tag{3}$$

Next, we feed the claim into the ERNIE model to get the claim representation $\boldsymbol{T_c}$. We also feed the evidence metadata e_m into the ERNIE model to get the evidence metadata representation $\boldsymbol{T_m}$.

$$\boldsymbol{T}_c = \text{ERNIE}(c) \tag{4}$$

$$\boldsymbol{T}_m = \text{ERNIE}(e_m) \tag{5}$$

We then add a relation-aware matrix to let the evidence detector sense the correlation between the target entity relation and the evidence metadata.

$$f_j = \boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{W}_f \boldsymbol{T}_m^{(j)} \tag{6}$$

where $\boldsymbol{T_c}, \boldsymbol{T}_m^{(j)}$ are the embedding representation of the claim and the j-th evidence metadata, respectively. $\boldsymbol{W}_f$ is $F \times F$ relation-aware matrix, which enables the detector to focus on the right location.

$$\alpha_j = \text{softmax}(f_j) = \frac{\exp(\boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{W}_f \boldsymbol{T}_m^{(j)})}{\sum_{k=1}^{N} \exp(\boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{W}_f \boldsymbol{T}_m^{(k)})} \tag{7}$$

Finally, we calculate relation-aware attention coefficient α_j for each evidence using the softmax function.

3.3.2 Medical Ontology-Enhanced Evidence Aggregator. When the evidence detector focuses on the accurate evidence, we then use a medical ontology-enhanced evidence aggregator to check the claim via reading comprehension of the multiple evidence. The objective of the evidence aggregator is to estimate the probability that the entity relation holds true given the retrieved evidence.

We first feed each claim-evidence pair (c, e_j) into ERNIE to get the representation of each claim-evidence pair $\boldsymbol{T}_v^{(j)}$.

$$\boldsymbol{T}_v^{(j)} = \text{ERNIE}(c, e_j) \tag{8}$$

Then, we incorporate the medical ontology graph $\mathcal{G}$ to get representation $\widetilde{\boldsymbol{T}}_v^{(j)}$, which can enhance the inference ability for textual entailment by utilizing medical ontology knowledge. The construction of $\mathcal{G}$ is described in 4.1.3.

$$\widetilde{\boldsymbol{T}}_v^{(j)} = \sum_{k \in \mathcal{N}_j} g_{jk} \boldsymbol{T}_v^{(k)} \tag{9}$$

where $g_{jk} \in \mathcal{G}$ indicates the ontology relation between the disease entity $e_d^{(j)}$ and $e_d^{(k)}$, $\mathcal{N}_j$ is set of neighbor nodes of $e_d^{(j)}$. The final hidden state representation $\boldsymbol{T_o}$ is obtained by gathering the multiple evidence information. The relation-aware attention coefficient $\boldsymbol{\alpha}$ is learned by the evidence detector in Eq. (7).

$$\boldsymbol{T_o} = \sum_{k=1}^{N} \alpha_k \widetilde{\boldsymbol{T}}_v^{(k)} \tag{10}$$

The final prediction $\text{Score}_{\text{msrv}}$ is calculated as follow:

$$\text{Score}_{\text{msrv}}^{(\text{i})} = \frac{\exp(\boldsymbol{W}_{out}^{\mathrm{T}} \boldsymbol{T}_o^{(i)})}{\sum_{j=1}^{C} \exp(\boldsymbol{W}_{out}^{\mathrm{T}} \boldsymbol{T}_o^{(j)})} \tag{11}$$

where $\boldsymbol{W}_{out} \in \mathbb{R}^{C \times F}$ denotes weight matrix, as before, F is the number of hidden dimensions of ERNIE, C is the number of prediction labels (here, $C = 2$). The final loss function is obtained as follows:

$$\mathcal{L} = -\frac{1}{m} \sum_{i} \hat{y}^{(i)} \log(\text{Score}_{\text{msrv}}^{(\text{i})}) + \lambda ||\theta||_2 \tag{12}$$

where θ denotes all trainable parameters, m is the number of training examples, $\hat{y}^{(i)}$ is the ground truth relation label for i-th example.

3.4 Interactive Collaborative-Training

In order to make the results of medical relation verification more explainable and better assist doctors to make a diagnosis, our deployed system is also aiming to improve the accuracy of the retrieved the evidence. However, it is a challenge to guarantee the accuracy of evidence due to the lack of evidence labels. Therefore, we introduce a variant of active learning method [22], i.e. interactive collaborative-training (ICT), to iteratively improve the accuracy of retrieved evidence through interactions with the MSRV model, evidence discriminator and medical experts. In our active learning settings, the informative instances should satisfy two properties: *uncertainty* and *importance*. The *uncertainty* means that the evidence discriminator cannot make confident predictions, and the *importance* means the evidence itself is important to MSRV for relation verification. The ICT is developed according to the above principle.

3.4.1 Evidence Discriminator. In order to improve the accuracy of retrieved evidence, we develop an evidence discriminator D_ϕ to score the retrieved evidence and then use it to filter out the wrong evidence. The evidence discriminator outputs a single scalar, which represents the confidence probability that whether the retrieved evidence can support the claim or not. If the evidence cannot support the claim, then it will be filtered and not involved the training of the MSRV model.

$$\text{Score}_{\text{e}} = \text{sigmoid}(D_\phi(\text{concat}(c, e))) \tag{13}$$

where c is the claim and e is the corresponding evidence, the evidence discriminator $D_\phi(\cdot)$ is built based on an ERNIE model.

3.4.2 Collaborative-Training Process. During training, two data pools are maintained and denoted as U (unlabeled data) and L (labeled data). Note that both U and L have golden labels for relation verification, while only L has golden labels for evidence. At each iteration, D_ϕ is trained on L, and the MSRV model M_θ is trained

Algorithm 1 One iteration of Interactive Collaborative-Training

Input: training sets U, L; evidence discriminator D_ϕ; base MSRV model M_θ; thresholds ϵ, δ; number of assigned evidence n;
Output: trained evidence discriminator D^*_ϕ; trained base model M^*_θ; updated training set U, L

```
Train D_φ on L; Train M_θ on U, L;
Initialize L' = ∅;
for each (claim, evidence) ∈ U do
    Acquire evidence score s_e via Eq. (13);
    Acquire relation-aware coefficient α via Eq. (7);
    if Entropy(s_e) ≥ ε then
        Add the (claim, evidence, α) to L'
        L' = sort(L', n); // sort L' and select top n
        Assign L' to medical experts for annotation;
    end if
    if Entropy(s_e) ≤ δ then
        Refine the evidence label of (claim, evidence) in U;
    end if
end for
L = L ∪ L', U = U\L'
return M*_θ, D*_φ, U, L
```

on U and L. After training, the D_ϕ makes evidence predictions on unlabeled instances. On the one hand, we first add the uncertain instances to L' and sort them with relation-aware coefficients calculated by Eq. (7), and then ask the medical experts to annotate the top-n evidence labels with the help of our system. On the other hand, we select the instances with high confidence, and we refine their evidence labels according to the output score of D_ϕ. After that, we get the updated U and L, which are used to update the D_ϕ and M_θ in the next iteration. The evidence with *negative* label is filtered and is not involved the training of the MSRV model M_θ in the next iteration.

In the first iteration (iteration 0), the initial labeled set L is a small set of instances already annotated by medical experts. The initial D_ϕ is trained on L. The initial U is the same as the experimental dataset for relation verification. The evidence of U is first acquired by document retriever and sentence selection module. The initial MSRV model M_θ is trained on U with relation labels. We iteratively process the above procedure, as the accuracy of the evidence meets the requirement (i.e. ≥ 0.8). The procedure of one iteration of ICT is described in Algorithm 1. The Entropy$(\cdot)$ means the entropy of probability distribution, the ϵ, δ are two thresholds.

4 EXPERIMENT

4.1 Data Description

4.1.1 Clinical Materials. We have collected a Chinese corpus containing 60 clinical medical textbooks (e.g. *Internal Medicine, Respiratory disease, Neurology, Pediatrics, Gastroenterology*, etc.) and medical encyclopedia data. In total, the corpus includes 8.72 million sentences and 23,722 diseases. The clinical materials are processed to the structured format, and then indexed by the ElasticSearch with the metadata. The metadata includes paths and titles (e.g. *clinical manifestation, examination, diagnosis, treatment, differential diagnosis, etiology, prevention, prognosis*, etc.). An example of a structured document is shown in Table 2.

4.1.2 Construct Corpus for Sentence Selection. In the synonym-aware sentence selection section, we first construct synonym pairs from de-identified Electronic Medical Records (EMRs) and the medical encyclopedia [2] according to the rule-based and the synthetic-based method. We construct 10k synonym pairs. Then we filter out the low-quality candidate pairs with pre-defined rules. We generate a synonymous entity-sentence set according to the selected synonym pairs. We retrieve the sentence containing the target entity or the synonym of the target entity as positive samples. As for negative samples, we randomly select a sentence in the same document at which the target entity locate. We take each entity-sentence pair as input and corresponding label as the prediction target. Our sentence selection module is then trained with the constructed corpus.

4.1.3 Medical Ontology Knowledge. Our medical ontology graph is built based on the Baidu Medical Knowledge Graph (Baidu MKG). Considering the efficiency in real-world applications, we transform the hierarchical structure data format into a flat adjacent matrix format and only keep the relations from nearest parent nodes or child nodes. It can avoid computing the complex hierarchical relations in real-time. The processed medical ontology graph $\mathcal{G}$ has 27,764 disease ontology relations and contains 16,492 disease entities.

4.1.4 Experimental Dataset for Relation Verification. Our relation verification dataset contains 32,823 disease-target entity relation pairs. The medical entity pairs are first extracted from de-identified Electronic Medical Records (EMRs) and online websites, and then annotated by the medical experts. As the main focus of this paper is to verify the medical entity relation, the disease-target entity relation extraction is regarded as the prepossessing step. To be more specific, we first utilize a constructed medical dictionary to extract medical entity mentions from the raw texts. Then we map the entity mentions to specific entity types, and the relation between the disease and target entity is inferred from the entity types. At last, we match entity pairs in the same text to possible knowledge triplets (i.e claim), the evidence of the claim is constructed by the sentence selection module. An example of the claim-evidence pair is shown in Table 1. We split the dataset into 3 parts: training, validation, and testing for our medical entity relation verification system. In Figure 2, we provide summary statistics of our experimental dataset.

4.2 Experimental Settings

For the word vector model, we set the length l_w of word vector to 200, initial learning rate α_w to 0.001, neighboring window size C_w to 5. For the document retrieval module, we retrieve the top k relevant document, where we set k is 10. For the sentence selection module, threshold τ_w of semantic similarity model and the threshold τ_s of synonym-aware model are set to 0.8. For both synonym-aware sentence selection module and MRC-based semantic relation verification module, we use the same network structure $ERNIE_{base}$ in all ERNIE fine-tuning tasks. The $ERNIE_{base}$ model has 12 layers, the hidden state dimension F is set to 768, the number of heads is set to 12. The learning rate α_e is set to 2e-5. For the synonym-aware sentence selection model, the maximum sequence length L^s_{max} is 128. For our MSRV model, we set the maximum sequence length

[2] http://www.a-hospital.com/

Method	Precision	Recall	F1-Score
FastText[2]	0.49	0.482	0.485
Synonym Model (SA)	0.653	0.725	0.687
GEAR[32]	0.740	0.701	0.719
KGAT[14]	0.792	0.714	0.750
MSRV (EA only)	0.815	0.700	0.753
MSRV (ED + EA)	0.841	0.697	0.762
MSRV (EA only + SA)	0.812	**0.822**	0.817
MSRV (ED + EA + SA)	**0.847**	0.815	**0.831**
MSRV (ED + EA + SA) + ICT	0.865	0.808	0.836

Table 4: Performance comparisons with the different baseline models on offline dataset. (threshold $\tau_s > 0.8$).

Method	Precision	Recall	F1-Score
MSRV (ED + EA + SA)	0.947	**0.768**	0.848
MSRV (ED + EA + SA) + ICT	**0.963**	0.762	**0.851**
Human	0.989	0.937	0.962

Table 5: Performance comparisons of Online Evaluation. (threshold $\tau_s > 0.8$).

L^v_{max} to 256. The batch size is set to 32. The L2 regularization parameter λ is set to 0.1. For ICT training, the network structure of the evidence discriminator D_ϕ is the same as $ERNIE_{base}$, the maximum sequence length L^d_{max} is set to 256. The log base of the entropy function is set to e, the ϵ and δ are set to 0.65 and 0.5, respectively. The number of assigned evidence n for each claim is set to 3, and we process the ICT for 2 iterations. We train our system with the paddlepaddle[3] deep learning framework. For the GEAR [32] and KGAT [14], we use the public source code and default parameter settings for the evaluation of the experiments.

4.3 Model Comparison

In this section, we compare our model with several baselines to verify the effectiveness of our approach.

- **FastText.** We obtain semantic similarity features via FastText [2], and only use the cosine distance as the probability of the relation prediction.
- **Synonym Model.** The synonym model is only using the result of the synonym-aware selection module as output. If any of the sentences are retrieved, the system will predict the claim being true.
- **GEAR.** We utilize the GEAR [32] model as one of the competitive baseline models, which achieves very promising performance in the general fact verification task.

[3]https://github.com/PaddlePaddle/Paddle

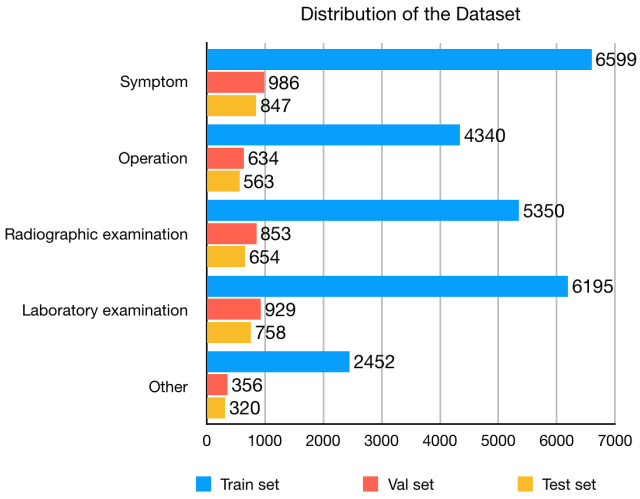

Figure 2: Summary Statistics of the Experimental Dataset

- **KGAT.** The KGAT [14] model use kernel-based attentions to improve performance, which is one of the state-of-art models in the general fact verification task.
- **MSRV (EA only).** MSRV is our fact verification model, EA means we only use medical ontology-enhanced evidence aggregator alone.
- **MSRV (ED + EA).** ED + EA means we use both relation-aware evidence detector and medical ontology-enhanced evidence aggregator.
- **MSRV (EA only + SA).** SA means we use the synonym-aware sentence selection module and MSRV (EA only).
- **MSRV (ED + EA + SA).** We utilize both relation-aware evidence detector, medical ontology-enhanced evidence aggregator and the synonym-aware sentence selection module for our relation verification.
- **MSRV (ED + EA + SA) + ICT.** After the MSRV (ED + EA + SA) training, we adopt the Interactive Collaborative-Training to improve the evidence quality.

4.4 Experimental Results

Table 4 shows the performance comparison with different methods on the test dataset. We observe that the FastText performs the worst, which is reasonable as it does not utilize any evidence with medical domain knowledge. The performance of the synonym model is better than FastText since it captures the synonym and implicit relation in sentences. The threshold of the synonym model τ_s is set to 0.8. The KGAT achieves better performance at a precision of 0.792 and a recall of 0.714 compared to GEAR. The GEAR and KGAT are very promising general frameworks for fact verification, however, due to without considering the practical challenges in the medical domain, both of them perform less effectively in this medical entity relation verification task. For our MSRV model, we set up two sets of comparative experiments. The first two models do not use synonym-aware sentence selection module. From table 4, we can see that the precision of the MSRV model with both relation-aware evidence

detector (ED) and medical ontology-enhanced evidence aggregator (EA) reaches 0.841, which is better than using an evidence aggregator alone. It means that our relation-aware evidence detector is useful for improving precision. As for the latter two experiments, MSRV (EA only + SA) utilizes a synonym-aware sentence selection module to achieve the recall of 0.822, which is higher than MSRV (EA only). It indicates that the SA module can help improve the recall. We observe that the MSRV (ED + EA + SA) performs the best precision of 0.847 and the best F1-score of 0.831, which proves that our proposed method is superior to other methods for medical relation verification. After the training of MSRV (ED + EA + SA), we adopt two iterations of ICT to improve the accuracy of retrieved evidence. The ICT method not only can improve evidence accuracy but also can improve the overall performance of the relation verification. Noting that for a fair comparison, we only compare the MSRV (ED + EA + SA) with other baseline models, as the procedure of ICT involved extra evidence label information. Two case study examples are shown in the appendix due to the space limit.

5 ONLINE EVALUATION

In this section, we show how our deployed medical verification system to solve the real-world medical verification task. We demonstrate the usability of our system by human evaluation from two perspectives: 1) the *effectiveness* to verify the relations and improve the accuracy of the retrieved evidence; 2) the *efficiency* to reduce human efforts.

Effectiveness. We randomly select 600 medical entity pairs from our medical knowledge graph (which has about 400k disease-target entity pairs) whose relations have been automatically verified by our system. Then we filter out the non-standard medical entities and finally obtain 546 candidate medical entity relation pairs for human evaluation. We employ three medical experts to manually evaluate each relation and the corresponding evidence, and label the correctness of relations and evidence by majority voting. Table 5 shows system performance on the online experiment. Our MSRV (ED + EA + SA) + ICT model can achieve the precision of 0.963 and the recall of 0.762. As shown in Figure 3, without ICT training, the initial evidence accuracy is 65.5%. After two iterations of ICT training, we get evidence accuracy of 82.5%, the improvement of evidence accuracy is significant. The row of *Human* in table 5 indicates the medical expert's performance evaluated by another senior expert, which is the upper bound for this verification task. The significance of our verification system is that we can automatically verify a huge number of medical entity relations with high precision, which can significantly reduce the time for medical experts.

Efficiency. Our system can also significantly reduce the time cost for manual verification of the medical knowledge graph. In some medical applications, human evaluation for every medical entity relations is necessary since high precision is required. For example, if the required precision is higher than the precision of our system (i.e. $\geq$ 0.98), human evaluation is necessary to manually verify every relation. We showcase the efficiency of our system to reduce the human effort by the following experiment.

Method	Efficiency (s/item)	Speedup
Human	106	-
Human + Our System	36	**2.94X**

Table 6: Performance comparisons on the efficiency of medical relation annotation.

The 546 candidate medical entity pairs are then assigned to two groups of medical experts for relation verification. Each group has 10 medical experts. Each medical expert is required to verify the medical entity relation through authoritative clinical materials and should at least retrieve one evidence to support the claim. For comparison, we only provide the system results to one expert group. The system results include the predicted probability and retrieved evidence. We then record the average time for experts to verify the candidate medical relations. As shown in Table 6, the average time for the expert assisted by our verification system is 36 seconds. In comparison, the average time for the other expert group is 106 seconds. Through this medical verification task, we can see that our system can significantly reduce the time of annotation by medical experts, which improves the efficiency of nearly 300%.

Additionally, in order to prove that our proposed ICT method can improve the efficiency of evidence annotation, we compare ICT with the Random Sampling strategy.

- **MSRV + RS.** At each iteration, we randomly select a number of instances and ask medical experts for annotation.
- **MSRV + ICT.** At each iteration, we select the most informative instances according to Algorithm 1 and ask the medical experts for annotation.

The annotation efficiency comparison between ICT and Random Sampling is shown in Figure 3. After initial training of our system, when adding the same number of labeled instances (i.e. 800), the ICT can achieve a much higher evidence accuracy (82.5% v.s 72.6%) compared to RS.

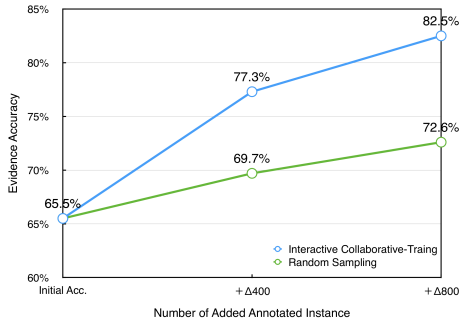

Figure 3: Annotation efficiency comparisons of Interactive Collaborative-Training and Random Sampling.

6 DEPLOYMENT

Figure 5 in Appendix Section A.2 gives an overview that how our proposed medical entity relation verification system applied on

Baidu CDSS, [4] which functions as a real-time professional assistant to doctors to guide them through standard diagnosis and treatment procedures, alerting them to potential errors and recommending suitable therapeutic plans. In this setting, the claim consists of the doctor's diagnosis, the medical entity (such as symptoms, signs, diagnosis, etc.), and the relation between the diagnosis and the medical entity. As shown on the left side of Figure 5, the doctor's diagnosis is *acute upper respiratory infection*, and below is the symptom entities which are automatically verified by our medical relation verification system. As shown on the right side of Figure 5, the disease entity below the *acute upper respiratory infection* is the differential diagnosis needed to be distinguished by the doctor, and the text shown in the orange box is the evidence retrieved by the synonym-aware sentence selection module and then evaluated by the evidence discriminator. With the assistance of our medical verification system, the doctor in the hospital can make a more reliable diagnosis with evidence retrieved from the authoritative clinical materials.

7 CONCLUSION

In this paper, we introduce a complete description of the implementation of our automatic medical entity relation verification system with large-scale machine reading comprehension. Our system is comprised of three modules: a document retrieval module, a synonym-aware sentence selection module, and an MRC-based semantic verification module. In addition, we introduce an interactive collaborative-training method to improve the evidence accuracy. The proposed synonym-aware sentence retrieval model retrieves the potential evidence that implicitly verifies the given claim. The MRC-based model contains a relation-aware evidence detector and a medical ontology-enhanced evidence aggretator to improve the precision of the relation verification module. We conduct extensive experiments on the offline dataset and applied our system for real-world medical entity relation verification tasks. The experiment results show that the performance of the proposed framework is superior to the other comparable models, and the verification system can significantly reduce the time for the medical expert verification task. To the best of our knowledge, it is the first deployed system to apply automatic medical relation verification techniques into real-world applications.

ACKNOWLEDGMENTS

Our work is supported by the National Key Research and Development Program of China No.2020AAA0109400.

REFERENCES

[1] Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In EMNLP. 534–545.
[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. In ACL. 135–146.
[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In ACL. 1870–1879.
[4] Jun Chen, Jingbo Zhou, Zhenhui Shi, Bin Fan, and Chengliang Luo. 2019. Knowledge abstraction matching for medical question answering. In BIBM. IEEE, 342–347.
[5] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In ACL. 1657–1668.

[4]The text in Figure 5 is translated from Chinese to English. An original Chinese version of Baidu CDSS is also shown in Appendix.

[6] Yang Deng, Yaliang Li, Ying Shen, Nan Du, Wei Fan, Min Yang, and Kai Lei. 2019. MedTruth: A Semi-supervised Approach to Discovering Knowledge Condition Information from Multi-Source Medical Data. In CIKM. 719–728.
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In ACL. 4171–4186.
[8] Hongliang Fei, Shulong Tan, and Ping Li. 2019. Hierarchical multi-task word embedding learning for synonym prediction. In KDD. 834–842.
[9] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In Fact Extraction and VERification Workshop. 103–108.
[10] Yu Hao, Xien Liu, Ji Wu, and Ping Lv. 2019. Exploiting Sentence Embedding for Medical Question Answering. In AAAI, Vol. 33. 938–945.
[11] Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking. In ACL. 8593–8606.
[12] Christopher Hidey and Mona Diab. 2018. Team SWEEPer: Joint Sentence Extraction and Fact Checking with Pointer Networks. In Fact Extraction and VERification (FEVER) Workshop. 150–155.
[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016).
[14] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In ACL. 7342–7351.
[15] Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. QED: A fact verification system for the FEVER shared task. In Fact Extraction and VERification (FEVER) Workshop. 156–160.
[16] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In AAAI, Vol. 33. 6859–6866.
[17] Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction. In ACL. 3916–3927.
[18] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In EMNLP. Association for Computational Linguistics, Austin, Texas, 2249–2255.
[19] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).
[20] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In EMNLP. 2383–2392.
[21] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In CIKM.
[22] Burr Settles. 2009. Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.
[23] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. AAAI (2020), 8968–8975.
[24] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In NAACL. 809–819.
[25] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification Shared Task. In Fact Extraction and VERification Workshop. 1–9.
[26] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In EMNLP Workshop. 353–355.
[27] Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative Adversarial Regularized Mutual Information Policy Gradient Framework for Automatic Diagnosis. In AAAI, Vol. 34. 1062–1069.
[28] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In NIPS. 5754–5764.
[29] Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Ucl machine reading group: Four factor framework for fact finding. In Fact Extraction and VERification (FEVER) Workshop. 97–102.
[30] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In AAAI. 5706–5713.
[31] Jingbo Zhou, Shan Gou, Renjun Hu, Dongxiang Zhang, Jin Xu, Airong Jiang, Ying Li, and Hui Xiong. 2019. A collaborative learning framework to tag refinement for points of interest. In KDD. 1752–1761.
[32] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In ACL. 892–901.

A APPENDIX

A.1 Case Study

Claim: (pulmonary cryptococcosis, **vomiting**, symptom)

Evidence1: Adverse reactions and countermeasures Adverse reactions of amphotericin B include chills, fever, headache, loss of appetite, nausea and **vomiting**, phlebitis, hypokalemia renal impairment, anemia and liver impairment, etc.
Meta data (title): Treatment
Probablity: 0.027 **Original Probablity**: 0.924

Evidence2: The adverse reactions of flucytosine include gastrointestinal reactions such as loss of appetite, nausea, **vomiting** and diarrhea, and bone marrow suppression of liver damage and rashes.
Meta data (title): Treatment
Probablity: 0.025 **Original Probablity**: 0.903

Claim: (pneumonia, **wet rales**, symptom)

Evidence1: The main clinical manifestations are cough, large amount of white foam mucus sputum, fever, physical examination of the lungs and smell of **wet rales**.
Meta data (title): Clinical Manifestation
Probablity: 0.967 **Original Probablity**: 0.953

Evidence2: People with secondary infections may have repeated fever, cough, audible lungs, and **wet rales**, similar to pneumonia. Sometimes hearing and tracheal tapping sounds may help diagnosis, but the diagnosis depends on bronchial fiber microscopy.
Meta data (title): Differential Diagnosis
Probablity: 0.051 **Original Probablity**: 0.842

Figure 4: Visualization of the results on our proposed verification framework.

In section 4.4, we provide a quantitative analysis of the experiment results. In this section, to help better understand that our MSRV model can better utilize the medical domain knowledge to achieve higher precision, we provide two cases from the test set.

Figure 4 shows two cases for our evaluation [5]. The first case is to verified the relation between *pulmonary cryptococcosis* and *vomit* via our verification system. As shown on the upside of Figure 4, two pieces of evidence are retrieved at the final stage. We compare the prediction of the MSRV (EA only) with MSRV (ED + EA). We found that the former model infers that two pieces of evidence are relevant with probability 0.924 and 0.903, respectively. However, the evidence is located at *treatment* paragraph, which actually said the side-effect of a drug, thus cannot support the relation between *pulmonary cryptococcosis* and *vomit*. The ED calculates the attention coefficient between *treatment* and the claim is 0.0775, and the final probability for evidence 1 is 0.027, for evidence 2 is 0.025, which correctly predicts the claim is not true.

The second case is to verify the relation between *pneumonia* and *wet rales*. As shown on the downside of Figure 4, without the relation-aware detector, the model retrieved two pieces of evidence. However, evidence 2 is not appropriate for verifying the given claim. The reason is that the retrieved evidence is under the paragraph of *differential diagnosis*, which actually describes the relation between the target entity and other disease entities. Our relation-aware evidence detector computes the attention coefficient score between *differential diagnosis* and the claim, and the final MSRV model will give little attention to this evidence.

[5]The text in Figure 4 is translated from Chinese to English.

A.2 System Interface

Figure 5 is a translated English version of Baidu CDSS. The original Chinese version of Baidu CDSS is shown in Figure 6. All medical entity relations displayed in the Baidu CDSS are validated by our verification system. The displayed evidence is first retrieved by synonym-aware sentence selection module and then evaluated by the evidence discriminator. With the assistance of our medical verification system, the doctor in the hospital can make a more reliable diagnosis with evidence retrieved from the authoritative clinical materials.

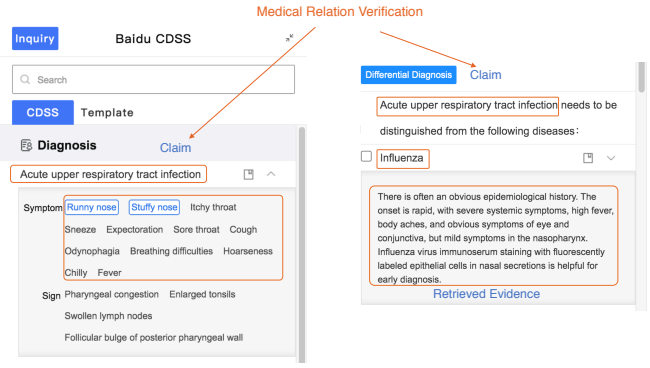

Figure 5: Medical Entity Verification Deployment on Baidu Clinical Decision Support System (CDSS).

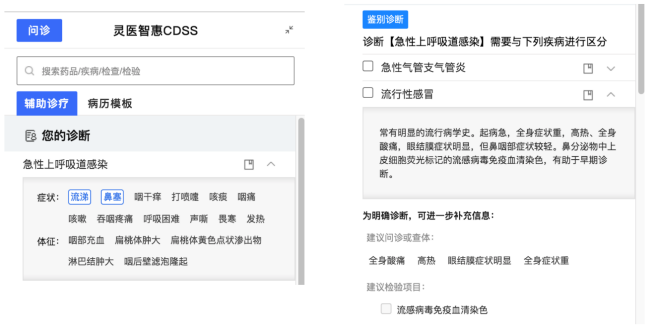

Figure 6: Medical Entity Verification Deployment on Baidu Clinical Decision Support System (Original Chinese Version)