Medical Term and Status Generation From Chinese Clinical Dialogue With Multi-Granularity Transformer

Mei Li, Lu Xiang, Xiaomian Kang, Yang Zhao, Yu Zhou, and Chengqing Zong, *Senior Member, IEEE*

***Abstract*—This paper describes a generative model for extracting medical terms and their status from Chinese medical dialogues. Notably, the extracted semantic information is particularly important to downstream tasks like automatic medical scribe and automatic diagnosis systems. However, how to effectively leverage dialogue context to generate medical terms and their corresponding status accurately remains less explored. Existing generative approaches treat dialogue text as a single continuous text, ignoring conversational characteristics like colloquialism, redundancy and interactions. Between the doctor and the patient, a variety of colloquial medical information is frequently discussed. Each speaker (doctor and patient) plays a specific role in the interaction's goals. As a result, the importance of role information and interactions between utterances cannot be overstated. Furthermore, existing generative approaches only use character-level tokens, disregarding word-level tokens, which are the shortest meaningful utterances in Chinese. In this paper, we propose a Multi-granularity Transformer (MGT) model to enhance the dialogue context understanding from multi-granularity features. We incorporate word-level information by adapting a Lattice-based encoder with our proposed relative position encoding method. We further propose a Role Access Controlled Attention (RaCa) mechanism for introducing utterance-level interaction information. Experimental results on two benchmark datasets illustrate our model's validity and effectiveness, achieving state-of-the-art performance on both datasets.**

***Index Terms*—Medical dialogue, multi-granularity, attention mechanism, natural language understanding, sequence to sequence learning.**

Manuscript received April 13, 2021; revised August 25, 2021 and October 14, 2021; accepted October 18, 2021. Date of publication October 26, 2021; date of current version November 6, 2021. This work was supported by the National Key R&D Program of China under Grant 2020AAA0108600. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Chengqing Zong.)*

Mei Li, Lu Xiang, Xiaomian Kang, and Yang Zhao are with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (CAS), Beijing 100049, China (e-mail: mei.li@nlpr.ia.ac.cn; lu.xiang@nlpr.ia.ac.cn; xiaomian.kang@nlpr.ia.ac.cn; yang.zhao@nlpr.ia.ac.cn).

Yu Zhou is with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China and also with Fanyu AI Laboratory, Beijing Fanyu Technology Company, Ltd., Beijing, China (e-mail: yzhou@nlpr.ia.ac.cn).

Chengqing Zong is with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, and with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: cqzong@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASLP.2021.3122301

I. Introduction

MEDICAL dialogue understanding (MDU), which intends to automatically detect medical terms and their status from doctor-patient dialogues, has recently received increased attention [1]–[6]. It plays an important role in task-oriented dialogue for automated diagnostic systems [7]–[10] and automated medical scribe for documenting clinical encounters [2], [11]. Notably, the conversation style data is challenging because the complex speech patterns of conversation lead to various medical information such as symptoms, drugs, and examinations, scattered in multiple dialogue turns. Moreover, the vocabulary can range from colloquial to medical jargon, which further increases the task's difficulty. Therefore, MDU typically includes entity detection, entity normalization and status inference, which aims to map colloquial entities or status expressions detected in dialogue into standard concepts. Fig. 1 shows a simple example, the expression "squeaking or snoring in the throat" is mapped into the medical term "gurgling with sputum". The status is used to indicate that whether a symptom is a patient suffering from or a physical examination has been done by the patient. In the example, the status of "gurgling with sputum" is "Uncertain," which can be inferred from the dialogue context.

For such a purpose, much work [11]–[14] has been proposed and has achieved promising results. These approaches can be grouped into two categories: classification-based methods and generative methods. In the classification-based methods [11]–[14], one way is to treat the task as a multi-label classification task. Another way [12]–[14] is to decompose the task into two stages, entity detection (the first stage), entity normalization and statue inference (the second stage). However, the classification-based methods have several limits: the two-stage method faces error accumulation and requires token-level annotation, the multi-label classification method treats each label independently, ignoring the relationship among entities. Differently, the generative methods [12] cast the task as a sequence generation problem, regarding standard medical terms and status as a target vocabulary, and generating terms and their status sequentially. It unifies entity recognition, normalization, and state inference. At the same time, sequence generation can take into account the associations among entities. In addition to the normalized final label, it does not require additional labeling information, such as token-level BIO (Begin-In-Out) labels. As a consequence, this approach is scalable across various

2329-9290 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

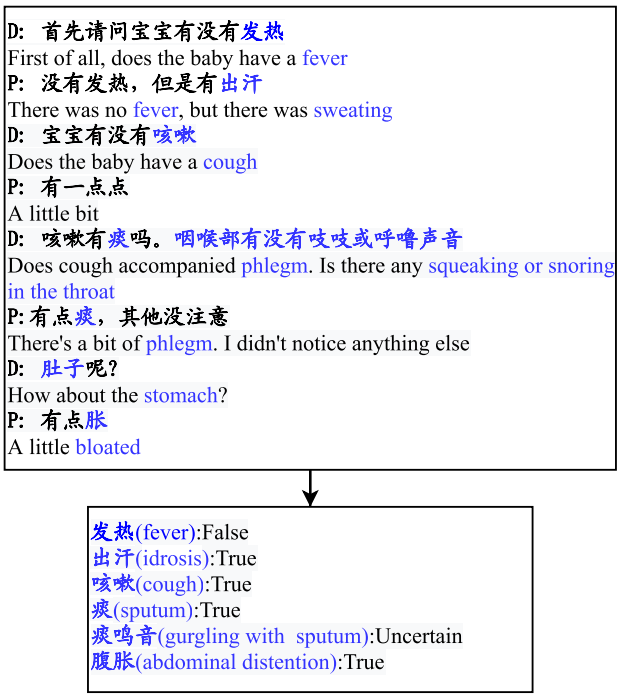


Fig. 1. A typical medical dialogue and the corresponding extracted medical terms and corresponding status. The "D" in front of utterances means doctor and "P" denotes patient.

medical specialties. Hence, the generative method is our preferred approach.

Although the current generative methods have many advantages, they treat the dialogue text as a long text. Still, they have some limitations: (1) Lack of word-level semantic. Most Chinese models take characters as the basic unit and learn representation, ignoring the semantics expressed in the word, which is the smallest meaningful utterance in Chinese. Due to the lack of explicit markers in fine-grained text to define the boundaries of words, it is often more difficult to identify entities and useful expressions. However, individual coarse-grained word-level representation learning often faces Out Of Vocabulary (OOV) and data-sparse problems [15]. (2) Lack of role information and interaction. The task-centric nature of conversation allows for effective communication of information by humans. The long text format prevents the systems from capturing dependencies over sentence boundaries and ignores the cross-turn interaction between speakers. We need to construct the interaction information between roles to make a better representation of the whole dialogue. In doctor-patient conversations, each of the speakers plays a specific role in the goals of the interaction. For example, the doctor is more likely to discuss medications and enquires about questions than the patient, as displayed in Fig. 1. Even the status tags of some tasks are directly related to roles [11], such as "doctor-positive," "patient-positive," and so on. Therefore, information about the role is also essential.

In this paper, we propose a new architecture named the Multi-Granularity Transformer (MGT) model to solve the above issues. (1) To reduce the problems caused by single granularity, we propose to use flat lattice-encoder [16] to combine word-level sequence with the character-level sequence. Each token of the two granularities is assigned a span of position indexes (head and tail) to model the distance between tokens. And we propose a new Relative Position Encoding (RPE) method to better model the distance between tokens. Through the comprehensive consideration of two granularities sequences, we can explicitly mitigate the deficiency caused by the single granularity. (2) Drawing inspiration from previous findings [17], we introduce a Role Access Controlled Attention (RaCa) mechanism. We add a special token (<psep> or <dsep>) at the beginning of each turn, which represents the roles of patient and doctor, respectively. We specify the access right of these tokens that can only access tokens in their current turn and role tokens in adjacent contexts. Thence these role tokens can aggregate the representation of the sequence and model the cross-turns interaction in a flat manner.

In a word, these modules contribute to obtaining better semantic dialogue representation from multi-granularity (character, word, and sentence). We compare our approach with existing models [11]–[13]. We conduct extensive experiments to verify the effectiveness of our proposed models on two published medical dialogue corpora. Experimental results demonstrate that our model can significantly outperform strong baselines. In summary, the contributions of this work are two-fold:

- We propose a Multi-Granularity Transformer (MGT) model to simultaneously capture the role-enhanced cross-turns interaction and integrate mixed granularity representations with a new relative position-coding module, thus fully model the dialogue text representation.
- We conduct extensive experiments and ablation studies on two datasets to examine the effectiveness of our MGT model. Experimental results show that our MGT model outperforms baselines and achieves state-of-the-art performance.

The rest of this paper is organized as follows. Section II presents the related work of medical dialogue and lattice-based methods. Section III introduces the background of the transformer method. Section IV introduces our proposed MGT method. The experimental setup is stated in Section V, and the experimental results are shown in Section VI. Section VII shows the conclusion.

II. Related Work

Currently, there is not much research work in the medical dialogue field. Most of the previous work has focused on extracting medical information and events from the patient-doctor conversation [1], [5]. This could involve extracting clinical entities and their properties [12], [13], [18], or medication regimen (dosage and frequency for medications) [19]. The extracted information can be used to generate a clinical note [2], [3], which contributes significantly to the efficiency of physicians in creating narrative reports [20]. It can also be used to construct automatic medical diagnosis systems [8]–[10], but this kind of information needs status inference, not just entity extraction. Among them, Finley *et al.* [2] proposes a pipeline method earlier. However, the details are scant. To alleviate the accumulation of errors caused by the pipeline, Finley *et al.* [3] proposes generating from an

Automatic Speech Recognition (ASR) transcript directly to a clinical note, but this performed poorly. After that, classification-based methods began to emerge. Du *et al.* [12], [13] and Lin *et al.* [14] propose two-stage methods. In the first stage, the entity is identified by a sequence tagging model. Based on this, two multi-layer perceptron classifiers are used in the second stage to perform entity normalization and entity status inference. However, these methods use the same framework and face several problems, including error accumulation and redundant annotation. On the other hand, Zhang *et al.* [11] uses a deep matching architecture to compute all candidates and all dialogue turns, considering the dialogue turn-interaction. Du *et al.* [12] also proposes using end-to-end generative models with pre-training, which achieves comparative results with classification-based methods. Prior work on this task is a source of SOTA modeling approaches that perform relatively well despite great challenges, while leaving much room for improvement.

Given recent advances in generative and pre-train models [17], [21], we explore the end-to-end paradigm for generative medical terms and status from patient-doctor conversations. End-to-end dialogue systems [22]–[24] and individual modules of dialogue system such as dialogue state tracking [25]–[28] and language understanding [29] have both made extensive use of generative techniques. Most previous work use single-granularity information, or use large-scale self-supervised pre-training models [23]–[25], [27], [29]. Using large-scale pre-training models, necessitates costly fine-tuning in the conversation domain [30]–[33], where the style differs significantly from the bulk of pre-training corpora. On the other hand, great experiments [16], [34]–[36] have shown that introducing extra lexicon can improve performance, in Chinese NER tasks. These work investigate lattice-structured LSTM [34], [37] or lattice-structure Transformer [16], [38]. In this paper, we extend the Transformer to handle multi-granularity information inspired by Li *et al.* and variants of Transformers [39]–[41]. It makes traditional Transformers learn character, word, and sentence information at the same time without changing the flat input. At the same time, each granularity can achieve effective interaction. This greatly improves the richness of features that the model can learn.

III. Background

We first introduce the transformer model [21], which is the backbone model of our method. The transformer encoder and decoder are composed of a stack of L transformer blocks. The l-th block takes a sequence of hidden representations $H^l = H_1^l, \ldots, H_n^l$ as the input and outputs an encoded sequence $H^{l+1} = H_1^{l+1}, \ldots, H_n^{l+1}$. Each layer of the Transformer encoder has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. A residual connection [42] around each sub-layer is utilized. In addition, each layer is followed by a normalization layer [43]. Similar to the encoder, the decoder inserts a third sub-layer, which performs multi-head attention over the outputs of the encoder.

A. Positional Encoding

Since the transformer model relies on a self-attention mechanism with no recurrence, the model is unaware of the sequential order of inputs. The Transformer injects absolute position embedding of the tokens into the sequence.

$$PE^{2i}(pos) = \sin\left(pos/10000^{2i/d_{\text{model}}}\right)$$
$$PE^{2i+1}(pos) = \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \quad (1)$$

where pos is the position and i is the dimension and d_{model} is the size of embedding.

$$H_i^1 = Emb\left[w_i\right] + PE[i] \quad (2)$$

where w_i denotes the i-th input token, Emb and PE denote a learned token embedding matrix and a positional embedding matrix, respectively. H^1 is the input of the first layer.

B. Masked Multi-Head Self-Attention

The first sub-layer of multi-head self-attention is modeled with multiple heads, computing independent self-attentional representations with individual parameters. The input consists of queries and keys of dimension d_k and values of dimension d_v. We denote queries, keys, and values as Q, K, and V, respectively. The output is a weighted sum of values. Scaled dot-product attention is employed to compute the attention weights. We calculate the outputs as:

$$\text{Attention}\ (Q, K, V) = \text{softmax}\left(\frac{A + M}{\sqrt{d_k}}\right) V \quad (3)$$

$$\text{A} = QK^T \quad (4)$$

where d_k denotes the dimension of key vectors. $\mathbf{M} \in \mathbb{R}^{m \times m}$ is an attention mask and m is the sequence length. Intuitively, if w_i is invisible to w_j, the $M_{i,j}$ will mask the attention score 0, which means w_i make no contribution to the hidden state of w_j.

Multi-head attention is to compute multiple independent attention heads in parallel and then concatenate the results.

$$\text{head}_i = \text{Attention}\left(H^l W_i^Q, H^l W_i^K, H^l W_i^V\right) \quad (5)$$

$$\text{MHA} = \text{ConCat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \quad (6)$$

where H^l denotes the input sequence of the l-th block, h denotes the number of heads, W_i^Q, W_i^K, W_i^V and W^O are trainable model parameters.

C. Position-Wise Feed-Forward Layer

The second sub-layer is a position-wise fully connected feed-forward network. And the residual connection is added around the two sublayers, followed by layer normalization. The output of the l-th block is calculated as:

$$O^l = \text{LayerNorm}(H^l + \text{MHA}(X^l)) \quad (7)$$

$$H^{l+1} = \text{LayerNorm}(\text{FFN}(O^l) + O^l) \quad (8)$$

where FFN is the position-wise feed-forward layer:

$$\text{FFN}(x) = \max\left(0, h_i W_1 + b_1\right) W_2 + b_2 \quad (9)$$

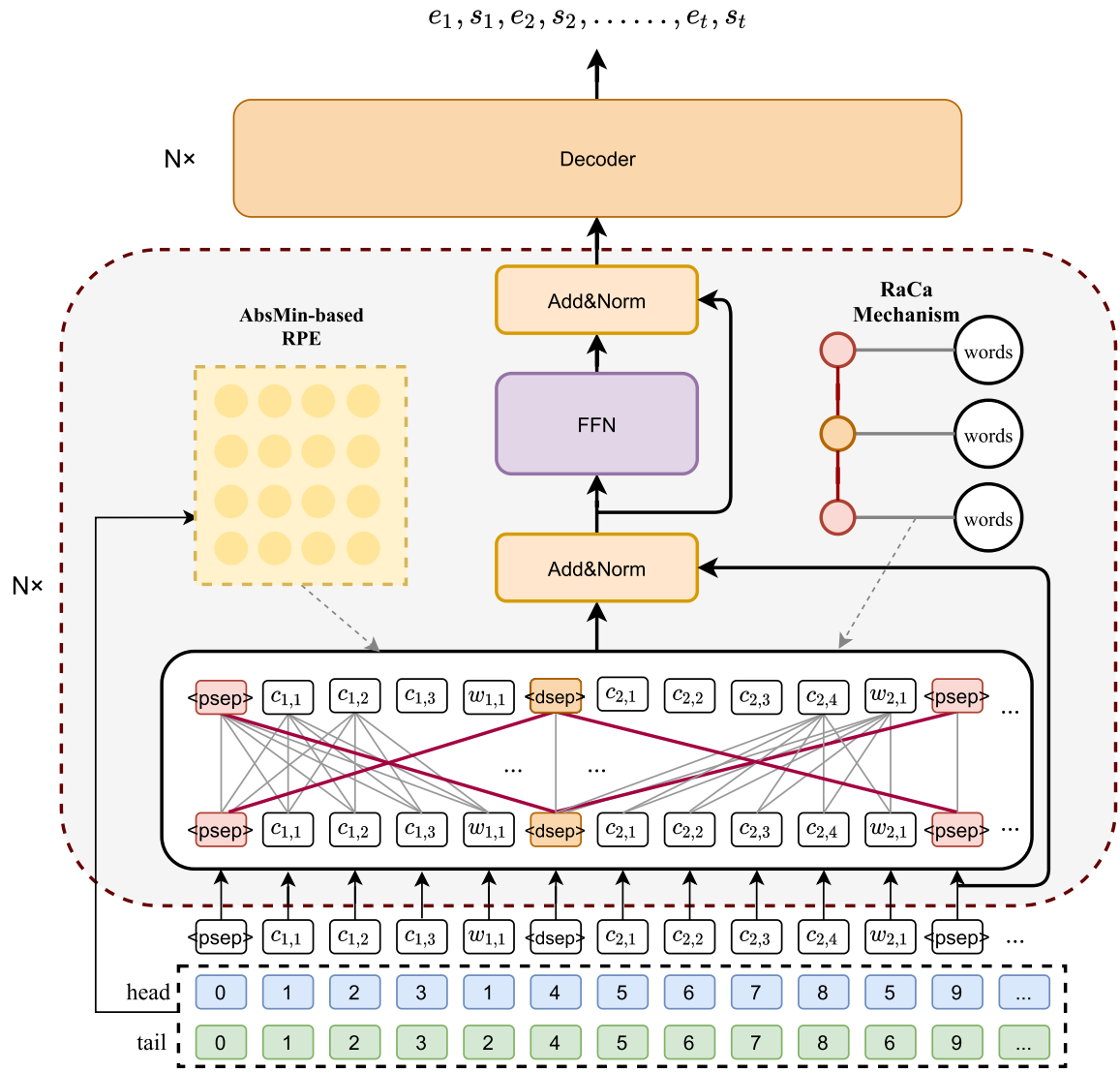


Fig. 2. Overview of our proposed MGT model. The input is a long concatenated sequence, and the position index of the head and tail is the continuous absolute position. The gray lines represent the full connection within the sentence. The red lines indicate the links between sentences. $c_{i,j}$ means character level tokens and $w_{i,j}$ means word-level tokens. e_i denotes entity and s_i denotes corresponding status.

where W_1 and W_2 are trainable parameters, b_1 and b_2 are parameter biases.

IV. Multi-Granularity Transformer

The goal of our model is to maximize the likelihood of a set of medical terms and status $Y = (e_1, s_1, \ldots, e_t, s_t)$, given a set of input dialogue sentences $X = (x_1, \ldots, x_m)$ as:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log \left(P_{\Theta} \left(\mathbf{Y} \mid \mathbf{X} \right) \right) \tag{10}$$

where e_i means the i-th medical term and s_i means the status of the i-th medical term, t means the number of medical terms in sample Y, m is the number of dialogue utterance, x_i means the i-th utterance, N is the number of training data.

A. Flat Lattice-Encoder

The Lattice-encoder is an effective and general way to represent multiple sub-sequences in a directed graph. In this paper, we use a flat lattice-encoder to model both character and word sequence simultaneously. It converts the inputs from the lattice into the flat structure by defining tokens as a set of spans. As shown in Fig. 2, each token (character or word) span includes a head position and a tail position, denoting the position index of the start character and end character in the original sequence. This method flatly represents lattice inputs while still maintaining the original structure.

1) Relative Position Encoding (RPE): Flat-lattice structures merge different granularity sub-sequences. Thus, the traditional absolute position encoding method described in section III-A is not suitable. Following Li *et al.* [16], we use relative position

between tokens based on four distances:

$$\begin{aligned} d_{i,j}^{hh} &= head_i - head_j \\ d_{i,j}^{ht} &= head_i - tail_j \\ d_{i,j}^{th} &= tail_i - head_j \\ d_{i,j}^{tt} &= tail_i - tail_j \end{aligned} \tag{11}$$

Previous work [16] uses a concatenated four-distance embedding following a nonlinear transformation as the position embeddings:

$$\begin{aligned} FourPos_{i,j} &= \mathrm{C}\,at(PE(d_{i,j}^{hh}), PE(d_{i,j}^{ht}), PE(d_{i,j}^{th}), PE(d_{i,j}^{tt})) \\ R_{ij} &= \mathrm{ReLU}\left(W_r\left(FourPos_{i,j}\right)\right) \end{aligned} \tag{12}$$

While this kind of relative position encoding is unfair to modeling the distance relationship between word spans. For example, for two words $token_i$ and $token_j$ with position head and tail pairs of [6, 8] and [3, 5], the distance between them is [3, 1, 5, 3] ($i \rightarrow j$). Although the two words are adjacent, the largest relative distance is still 5 and the most frequent distance is 3. Even the relative distance to oneself is not zero ($i \rightarrow i$ is [0, -2, 2, 0])

Therefore, we propose a new relative position encoding, which can better capture the distance between two types of tokens and be aware of the directionality at the same time. We only keep one distance that has the smallest absolute value. In the preceding example, the distance between $i \rightarrow j$ is 1, the distance between $i \rightarrow i$ is -1, and the distance between $i \rightarrow i$ is 0. Therefore, it can distinguish different directions and distances. The formula for calculating our Relative Position Encoding (RPE) is as follows:

$$\mathrm{pos}_{i,j} = \mathrm{AbsMin}([d_{i,j}^{hh}, d_{i,j}^{ht}, d_{i,j}^{th}, d_{i,j}^{tt}]) \tag{13}$$

where AbsMin is a function that keeps the element with the smallest absolute value in four distances. For example, $\mathrm{AbsMin}([2, -1, 3, 2]) = -1$.

We then calculate the distance embedding by using the equation (1). The final relative position encoding is a simple nonlinear transformation of the distance embedding:

$$R_{ij} = \mathrm{ReLU}\left(W_r\left(PE(pos_{i,j})\right)\right) \tag{14}$$

where R is a matrix and $R_{i,j}$ indicates the relative distance embedding between i-th token and $j-th$ token.

2) Lattice-Aware Self-Attention: Then the self-attention with relative position encoding [44] is calculated as:

$$\begin{aligned} \mathbf{A}_{i,j}^{rel} = {} & \mathbf{W}_q^\top \mathbf{E}_{x_i}^\top \mathbf{E}_{x_j} \mathbf{W}_{k,E} + \mathbf{W}_q^\top \mathbf{E}_{x_i}^\top \mathbf{R}_{i,j} \mathbf{W}_{k,R} \\ & + \mathbf{u}^\top \mathbf{E}_{x_j} \mathbf{W}_{k,E} + \mathbf{v}^\top \mathbf{R}_{i,j} \mathbf{W}_{k,R} \end{aligned} \tag{15}$$

where $\mathbf{W}_q, \mathbf{W}_{k,R}, \mathbf{W}_{k,E} \in \mathbb{R}^{d_{model} * d_{head}}$ and $u, v \in \mathbb{R}^{d_{head}}$ are learnable parameters, d_{model} and d_{head} denote the dimension of hidden state and each head.

The multi-head lattice attention is calculated as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{A^{rel} + M^r}{\sqrt{d_k}}\right) V \tag{16}$$

where the A^{rel} matrix is calculated by (15). M^r is the mask matrix, and we will discuss it later. The following steps after calculating the $head_i$ are the same with vanilla Transformer.

B. Role Access Controlled Attention (RaCa)

The Lattice-encoder uses a self-attention mechanism similar to the vanilla Transformer with fully connected attention connections. Then treating dialogue as a long sentence will lead to difficulty to capture turn-level information. However, common hierarchical mechanisms need to add additional parameters. Thus, we propose the RaCa model, which models the hierarchical structure within the long sentence without increasing any parameters. In each turn, we add a special token <psep> or <dsep>, which represents the roles of patient and doctor, respectively. The input sequences concatenate several turns, and each role token can interact with two categories of tokens. The first category is the tokens in the same turn. Previous experiments [17] have shown that a special classification token ([CLS]) in front of every example can be used as the aggregate sequence representation. Therefore, the role token can collect the relevant information from the rest of the turn tokens. The second category is the role tokens from the previous and subsequent turn. Then the role token not only models turn-level information but can also capture context information. Conversely, the token in the sentence can interact with the sentence representation containing context information. Formally, we obtain masks as follows:

$$m_{ij} = \begin{cases} 0 & \text{if } i \in \mathcal{S}(j) \vee j \in \mathcal{N}(i) \\ -\infty & \text{else} \end{cases} \tag{17}$$

where $\mathcal{S}(j)$ means all the tokens of the sentence in which the j-th token is located, $\mathcal{N}(i)$ denotes the neighborhood of sentence role tokens. The above mask is designed to be plugged into each Transformer layer via the masking term M^r in (16).

C. Decoder

Except for the mask in cross attention, the decoder in our model is the same as the decoder in the vanilla Transformer. There are some duplicate tokens in our context since it comprises both character-level and word-level tokens. With the premise that coarse-grained information is simpler for feature selection, we mask the redundant character-level tokens and retain word-level tokens in the decoder stage.

V. Experiments

A. Datasets and Evaluation Metrics

Chunyu Dataset [11]: The corpus includes 1,120 dialogues collected from a Chinese medical consultation website, ChunyuDoctor[1], under the topic of cardiology. This website archives doctor-patient consultation discussions, including chat content and role information, and conceals users' personal information. The four primary term categories defined in this corpus are symptom, surgery, test, and other info. The terms quantity of

[1][Online]. Available: https://www.chunyuyisheng.com

TABLE I
DESCRIPTION THE STATISTICS OF TURNS AND TOKENS ON TWO DATASETS

Description	Chunyu	CMDD
Train(window)	12931	30333
Dev(window)	2587	9806
Test(window)	2694	9866
Term number	71	146
Status number	5	3
#Turn/Conversation	16.16	26.50
#Medical Term/Conversation	8.42	15.14
#Token/Conversation	369.71	613.45
#Token/Turn	22.87	23.15

symptom, surgery, test, and other info, respectively, are 45, 4, 16, and 6. The term statuses are defined as follow: Patient-pos, Patient-neg, Doctor-pos, Doctor-neg, and Unknown. The term's triple form is as follows: (symptom, cold, Patient-pos). The annotation is a window-to-information approach, which implies that the information inside a segment of conversation is directly annotated without word-level sequence labeling information, and the window size is 5. The total number of windows is 18,212. The corpus is divided into train/dev/test sets of size 800/160/160.

CMDD Dataset [14]: This corpus is constructed from the pediatric department of a Chinese online health community,[2] which contains 2,067 conversations, focusing on four diseases, "upper respiratory infection," "functional dyspepsia," "infantile diarrhea" and "bronchitis". The CMDD corpus only contains symptoms. The symptom status is divided into three categories: True, False, and Uncertain. This corpus applies a character-level sequence labeling schema. Each symptom-related character is annotated by the BIO label. The total conversations are split by 3:1:1 to obtain a train/dev/test set. The original data of multiple sentences of the same role are treated as multiple turns. In our experiment, we merged continuous multiple sentences from a role into one turn. In this paper, we also segment the dialogues into windows, and the window size is 5. Table I shows more detailed statistics about the two corpora.

We segment all dialogue sentences using jieba.[3] To optimize the preservation of term integrity, we specify our own lexicon, which comprises a significant number of medical words. In the original Chunyu dataset, several statuses for the same medical term would be displayed in one window, and we only preserve the final state that is suitable for the conversation system in our experiment. In the CMDD dataset, we only removed the repeating medical term-status pairs in generative models.

Evaluation Metrics: We evaluate the results of each segmented window and show the micro-average of all the test windows. Following Zhang *et al.* [11], we consider two main metrics. **Full**: only the category (Only Chunyu corpus), medical term, and their corresponding status are strictly correct; the triples or tuples are considered valid. **Term**: only consider medical terms, regardless of status. It is worth noting that not all windows have golden labels; sometimes they are empty. In this case, if the prediction is also empty, then we set Precision, Recall, and F1-score to 1, otherwise 0.

[2][Online]. Available: https://www.muzhi.baidu.com
[3][Online]. Available: https://github.com/fxsjy/jieba

TABLE II
MAIN HYPER-PARAMETERS FOR TRANSFORMER-BASED AND BERT-BASED MODELS

Model	Transformer-based	BERT-based
enc/(dec) layers	3/3	12
hidden size	256	768
ffn size	507	3072
head num/size	4/64	12/64
dropout	0.1	0.1
learning rate	3e-4	3e-5

B. Baselines

We compared our proposed model MGT against three different type models:

- *SAT [12]:* a two-stage sequence labeling based method. The entity span is identified by the sequence tagging model(Transformer+CRF) using a generic tag set (BIO) in the first stage. Based on the hidden representation of identified spans, two multi-layer perceptron classifiers are used in the second stage to normalize the entity and infer their status.
- *FLAT-SAT*: we also employ the state-of-art sequence labeling model FLAT [16] in two-stage method, in which FLAT recognizes entities in the first state and then classifiers normalize entities and infer their status.
- *BERT-SAT & BERT-SAT$_f$:* two-stage methods with BERT encoders. And BERT-SAT$_f$ denotes the model fine-tuned by our training data.
- *MIE [11]:* a multi-label classification method based on a deep matching model that can make use of the interaction between dialogue turns.
- *Hier-Transformer:* A character-level Transformer model with hierarchical encoder.
- *Transformer-char:* A normal character-level Transformer model, treating the dialogue segment input as a long sentence.
- *Transformer-word:* A normal word-level Transformer model with the same structure as the Transformer-char model.

C. Implementation Details

The MIE model uses pre-trained 300-dimensional Skip-Gram [45] embeddings, the hidden size of Bi-LSTM is 400 and the drop rate is 0.2. All hyper-parameters are the same as described in Zhang *et al.* [11]. Transformer-based models and BERT-based models, including SAT and all generative models, use the hyper-parameters as described in Table II. The beam size for generative models is set to 5. The classifiers used in SAT-based methods are four layer multi-layer perceptrons. We use the Adam [46] optimizer to train the models. The hyper-parameters of the Adam optimizer: we set two momentum parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The Transformer encoder

TABLE III
TERM METRIC RESULTS OF DIFFERENT MODELS BOTH ON CHUNYU DATASET AND CMDD DATASET

Model	Chunyu			CMDD		
	Precision	Recall	F1-score	Precision	Recall	F1-score
MIE	0.913	0.815	0.844	0.867	0.831	0.835
SAT	-	-	-	0.885	0.878	0.872
FLAT-SAT	-	-	-	0.893	0.888	0.880
BERT-SAT	-	-	-	0.853	0.812	0.817
BERT-SAT$_f$	-	-	-	0.907	0.896	0.892
Hier-Transformer	0.899	0.886	0.881	0.885	0.856	0.861
Transformer-Char	0.922	0.903	0.903	0.901	0.885	0.887
Transformer-Word	0.906	0.878	0.880	0.848	0.796	0.807
Lattice-Transformer	0.913	0.894	0.893	0.921	0.901	0.904
RaCa-Char-w/o role	0.926	0.904	0.905	0.905	0.886	0.888
RaCa-Char	0.926	**0.905**	**0.906**	0.910	0.887	0.890
RaCa-Word	0.917	0.856	0.873	0.840	0.774	0.795
MGT	**0.938**	0.886	0.901	**0.928**	**0.902**	**0.907**

TABLE IV
FULL METRIC RESULTS OF DIFFERENT MODELS BOTH ON CHUNYU DATASET AND CMDD DATASET

Model	Chunyu			CMDD		
	Precision	Recall	F1-score	Precision	Recall	F1-score
MIE	0.718	0.689	0.687	0.747	0.733	0.727
SAT	-	-	-	0.751	0.753	0.742
FLAT-SAT	-	-	-	0.717	0.729	0.712
BERT-SAT	-	-	-	0.656	0.623	0.626
BERT-SAT$_f$	-	-	-	0.752	0.752	0.742
Hier-Transformer	0.707	0.699	0.695	0.746	0.721	0.725
Transformer-Char	0.693	0.682	0.680	0.743	0.720	0.724
Transformer-Word	0.669	0.654	0.652	0.702	0.656	0.666
Lattice-Transformer	0.707	0.699	0.695	0.784	0.759	0.764
RaCa-Char-w/o role	0.722	0.710	0.709	0.754	0.733	0.736
RaCa-Char	0.732	**0.720**	0.718	0.765	0.743	0.747
RaCa-Word	0.688	0.660	0.665	0.719	0.657	0.675
MGT	**0.753**†	0.717†	**0.727**†	**0.800**†	**0.768**†	**0.776**†

and Lattice-based Transformer encoder are all pre-trained with masked language modeling objective [17] on the mix of two corpus sentences. The mask rate is 15% which is following previous work [17].

VI. RESULTS AND ANALYSIS

In this section, we compare our model with existing studies on the experimental datasets. Table IV and Table III detail the results of our experiments. Since the Chunyu dataset has no token-level label information, we do not test the SAT method on the Chunyu dataset. To evaluate the contribution of individual components of our model, we also experiment with model variants where some of the components are removed or added. In detail, Lattice-Transformer denotes that only keep Lattice-encoder in MGT, which only takes character and word granularity into account. The RaCa model denotes that our RaCa module is added to the vanilla Transformer model with character-level inputs (RaCa-Char) and word-level inputs (RaCa-Word). Furthermore, "RaCa-Char-w/o role" means we replace two special role tokens with the same token in the experiment of the RaCa-Char model to verify the effect of the role information.

A. Main Results

From Table IV we can see that our MGT model achieves an F1 score of 72.7% on the Chunyu corpus and 77.6% on the CMDD corpus, which is considerable improvement in the real and complex medical dialogue. On the Chunyu dataset, our model outperforms the previous MIE results by a large margin of 4.0% and 4.9% on the CMDD dataset. Our model outperforms the SAT model by 3.4%, which demonstrates the power of integrating multi-granularity information. And the (†) indicates that the improvement over baselines is statistically significant where

$p < 0.05$. Besides, from the results of the ablation experiment, we can see that all the proposed components are effective in contributing the final performance as expected.

Comparison with classification-based models: As shown in Table IV, the SAT model is a strong baseline. The BERT-SAT is improved greatly after fine tuning, demonstrating that the domain of a large-scale pre-training model is considerably distinct from the area of medical conversation. However, our model outperforms strong baselines without the use of extra corpus, indicating that field knowledge (word segmentation based on medical dictionaries and incorporating conversation interaction) may significantly increase the model's performance and data utilization rate. The FLAT-SAT model provides word information on the basis of the SAT model, which enhances the SAT model's performance on the Term metric but not on the Full metric. This demonstrates that merely incorporating words into characters does not assist the model in carrying out long-distance semantic dependency and interaction between dialogue turns, both of which are critical for state inference. In our MGT model, we mainly incorporate character information into words. Our approach of relative position encoding encodes word-level information more fairly. And we employ coarse-grained word-level information and sentence level interaction information in the generating process. This enables the model to capture long-distance semantic dependent information more effectively.

Comparison with generative models: From Table IV we can see that the word-level Transformer model does not give better performance than the char-level Transformer. Moreover, they drop by more than 2% on the Chunyu dataset and more than 6% on the CMDD dataset, indicating that the word-level model may suffer from sparse data problems. However, we find that the performance of the multi-granularity method MGT is significantly better than individual granularity methods, which verifies the effectiveness of the proposed multi-granularity network for medical dialogue understanding. Moreover, our RaCa-Char model surpasses Hier-Transformer on both datasets without introducing additional training parameters, demonstrating the effectiveness and simplicity of our RaCa mechanism. The vanilla Transformer-Char model achieves good performance on the Term metric, especially on the Chunyu dataset, but it is worse than other baselines on the Full metric. This indicates that it lacks modeling of sentence semantics and does not have enough ability to infer the status. The Hier-Transformer model results are opposite, which suggests that term detection is more dependent on local information. In contrast, our MGT model is better than Transformer-Char on Full metric but achieves comparable results on Term metric, which shows that our model has better inference ability without losing the term detection ability. We also find that our MGT model has a noticeable improvement on the Term metric on the CMDD dataset. One possible reason for our analysis is that the Chunyu data set has fewer term categories.

B. Model Analysis

Ablation Studies: Comparing Lattice-Transformer with the Transformer-Char and Transformer-Word models, as suggested in Table IV, we can see that the mixture of characters and words is better than any single-granularity model. This indicates mutual complementation in the combination of char and word level information can contribute to a more comprehensive understanding of dialogue semantic. Comparing the RaCa (Char and Word) models with Transformers (Char and Word) models, we can see significant improvements. It shows the importance of cross-turn interaction and demonstrates that the RaCa module benefits medical term and status generation. Compared to the "RaCa-Char-w/o role" model with the "RaCa-Char" model, we can see obvious drops in performance, especially on the Chunyu dataset, whose status is directly related to roles. This illustrates the significance of considering the role information. Furthermore, we also find that the Lattice-Transformer produces better improvements than RaCa on the CMDD dataset. However, the opposite result is obtained on the Chunyu dataset. This suggests that both modules work but behave differently on different datasets.

TABLE V
RESULTS ON FULL METRIC OF TWO RELATIVE POSITION ENCODING METHODS

Dataset	Method	Precision	Recall	F1-score
Chunyu	AbsMin	**0.707**	**0.699**	**0.695**
	FourPos	0.686	0.667	0.669
CMDD	AbsMin	**0.784**	**0.759**	**0.764**
	FourPos	0.770	0.744	0.750

Effect of Relative Position Encoding: We studies the effects of different relative position encoding methods on two datasets, and the results are shown in Table V. We explored two relative position encoding settings in Lattice-Transformer: "AbsMin" indicates the method we proposed in this paper, "FourPos" indicates a non-linear transformation of the four distances proposed in reference [16]. We see that the method we proposed works better than the "FourPos" method on both datasets. In reference [16], four distances are used in the model, introducing much redundant position information. Meanwhile, our proposed method is more concise and helps the model learn useful information.

Effect of Different Interactions: According to the performance changes in the ablation models, we find that the use of the RaCa module can significantly improve the performance. In our RaCa module, we consider the hierarchical dialogue structure modeling the interaction among turn levels. However, there is a question: is there any other structure in dialogue text better than hierarchical structure? Therefore, we conduct experiments with different interaction structures, as shown in Fig. 3. Following the intuitive characteristics of the dialogue below: 1) Two adjacent sentences have the closest relationship. The "adjacent" structure means that the role token has access to both the current utterance and the previous utterance, modeling the adjacent relationship of dialogue. 2) Current utterance is always based on the overall meaning of the previous utterance. The "interlace" structure means that current utterance tokens have access to both current utterance tokens and previous utterance level representation token. The "hierarchical" structure denotes our original RaCa module. The results are shown in Table VI. It can be concluded

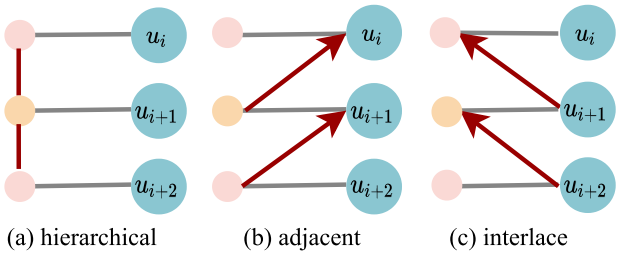


Fig. 3. Overview of different dialogue structure. "u_i" means all tokens in utterance i. The pink and orange node denote the different role tokens.

TABLE VI
EXPERIMENT RESULTS WITH DIFFERENT STRUCTURE ON CMDD DATASET

Dataset	Method	Precision	Recall	F1-score
CMDD	hierarchical	**0.800**	**0.768**	**0.776**
	adjacent	0.782	0.761	0.765
	interlace	0.781	0.758	0.762
Chunyu	hierarchical	**0.753**	**0.717**	**0.723**
	adjacent	0.749	0.704	0.717
	interlace	0.722	0.702	0.706

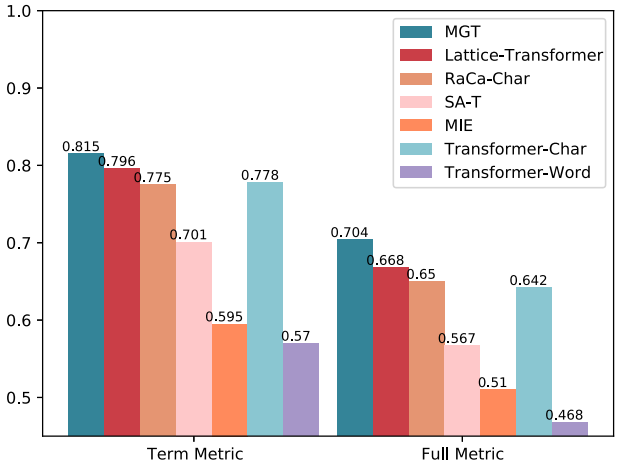


Fig. 4. The F1 score of different models in the "Implicit Terms" scenario.

from the table that the hierarchical structure is the most stable and effective.

C. Qualitative Analysis

Furthermore, to verify the effect of the model in complex dialogue scenarios, we randomly screened out the following three types of data from the test set:

I) Implicit terms, which denotes standard terms in the reference or their common expression, for example "腹泻 (diarrhea)" and its common expressions, such as "拉肚子 (diarrhea)" and so on, do not explicitly appear in the dialogue. The F1-score results on the Term metric and the Full metric are shown in Fig. 4. From the results, we can see that classification-based methods (SAT and MIE) drop by a large number of points. Especially the MIE model, which uses the standard term representation to match the dialogue context, drops sharply when there is a big difference between terms and context. In contrast, the performance of generative models is better in this case and our full model MGT performs best. This indicates that the generative model has advantages in the case of complex term expressions. After adding word-level information, there is a certain improvement both on the term metric and the full metric, which denotes that word-level information is useful for inferencing this kind of flexible expression.

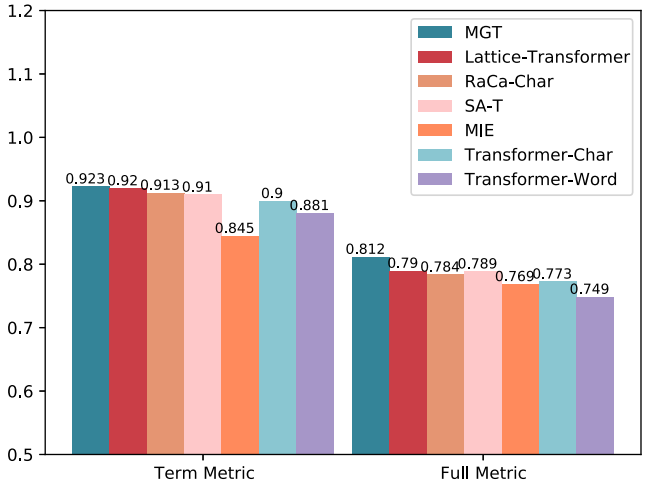


Fig. 5. The F1 score of different models in the "Implicit Status" scenario.

II) Implicit status, which denotes common status expressions such as "Not have" or "Yes" etc., do not exist in dialogue. This is a common situation in conversations as status information is usually used as an invisible premise and needs to be inferred from context. For example "What to do with enteritis?, " the status of "enteritis" is "True" by default, but in "What should I do if it is enteritis?, " the status of "enteritis" is "Uncertain". The results are shown in Fig. 5. From the figure, we can see that models generally perform well. The SAT model, which can locate the positions of terms, has an advantage in this scenario. However, our Lattice-Transformer and RaCa-Char model all achieve comparable results without intermediate annotation information and better than vanilla Transformer. This indicates that adding coarse-grained information (word or turn) can enhance the model's ability to understand the high-level semantic.

III) Interaction. Another common scenario is that the status of the term needs to be inferred through the interaction of speakers. We filter out these terms and calculate the accuracy of these terms with different models. Sometimes the answer may not appear immediately after the question, and there is some other content inserted in the font of the answer. The farther the distance, the more difficult it is to infer the status. Therefore, we classify terms according to the distance between the answer and the question. The results are shown in Fig. 6. "d = 0" means that the answer appears immediately after the question. "$1 \leq$ d $\leq$ 5" means there are no more than five characters between the answer and the question. From the results, we can see that the turn level information will be more effective as the distance increases. The MIE is better than the vanilla Transformer, and the RaCa-Char model surpasses the Lattice-Transformer model as

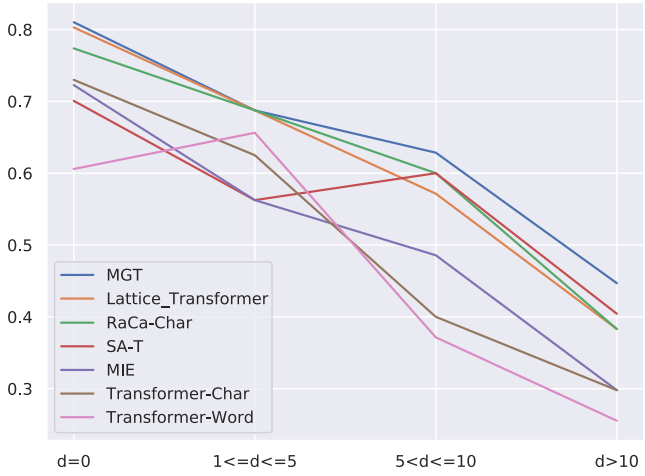


Fig. 6. The accuracy of different models in the "Interaction" scenario.

the distance increases. The SAT model can locate terms and still has advantages in this scenario. But our full model can process multi-granularity features simultaneously and has a stronger ability to understand different types of dialogues.

Case Study: We also present a qualitative analysis based on the case study. Fig. 7 displays the results produced by different models. As we can see, our MGT model generates more accurate medical terms and status for the doctor-patient conversation compared to baselines. All baselines ignore the symptom "gurgling with sputum (痰鸣音)," which is described more obscure compared to "fever" and "cough". While our model produces "gurgling with sputum (痰鸣音)" and corresponding status correctly, showing its efficacy of capturing global and local semantic information. We also found that our model has better status reasoning ability and the ability to capture relevant information. In example II, the status needs to be inferred through the interaction and the semantic information of the sentence. In this case, the MIE model fails to detect the state information, the Transformer-Char trends to generate three consistent status even though it only has one error. Our model, on the other hand, can correctly detect the corresponding entity and status information.

Efficiency of MGT: We also compare the running times of several models. Fig. 8 shows the running time of different models compared to the vanilla Transformer model. As seen in the figure, the Transformer, which benefits from parallelism, is quicker than the LSTMs, and the generative models do not fall short of the two-stage model. The speed of the MGT drops in comparison to the Transformer, SAT and BERT-SAT models, although it still offers benefits over other baselines. In order to illustrate the generalization of our model, we also compare our model results with different hyper-parameters. The results are shown in Table VII. The first line is our original reported results. The experimental results reveal that the model is relatively stable, with the F1 value of the modified parameter model remaining greater than 0.77.

Visualization of attention: For better understanding, we visualize the context attention distributions of the MGT model in Fig. 9 when generating medical terms and their status. From the

Example I (Implicit Term)	**Results**	
D: 首先请问宝宝有没有发热 First of all, does the baby have a fever P: 没有发热，但是有出汗 Therefore was no fever, but there was sweating D: 宝宝有没有咳嗽 Does the baby have a cough P: 有一点点 A little bit D: 咳嗽有痰吗。咽喉部有没有吱吱或呼噜声音 Does cough accompanied phlegm. Is there any squeaking or snoring in the throat P: 有点痰，其他没注意 There's a bit of phlegm. I didn't notice anything else	**SA-T**	发热(fever):False 发热(fever):False 出汗(idrosis):True 咳嗽(cough):True 咳嗽(cough):True 痰(sputum):True
	MIE	发热(fever):False 发热 (fever):Uncertain 出汗(idrosis):True 咳嗽(cough):True 痰(sputum):True
	Trans-Char	发热(fever):False 出汗(idrosis):True 咳嗽(cough):True 痰(sputum):True
Ground Truth: 发热(fever):False 出汗(idrosis):True 咳嗽(cough):True 痰(sputum):True 痰鸣音(gurgling with sputum):Uncertain	**MGT**	发热(fever):False 出汗 (idrosis): True 咳嗽(cough):True 痰(sputum):True 痰鸣音(gurgling with putum): Uncertain
Example II (Interaction)	**Results**	
D: 医生您好。半年了 Hello, doctor. It's been half a year P: 做过24小时动态心电图吗。得过心肌炎吗 Have you had a 24-hour Holter ECG. Have you ever had myocarditis? D: 这个就是24小时动态的呀。给您的照片。没有 This is a 24-hour Holter ECG. The photo for you. Nothing P: 最近感冒来吗 Did you catch a cold recently? D: 也没有 Nothing either	**MIE**	感冒(cold):Uncertain 心肌炎 (myocarditis) : Uncertain
	Trans-Char	感冒(cold):P-neg 心肌炎(myocarditis) : P-neg 心电图(ECG):P-neg 发绀(Cyanosis):P-neg
Ground Truth: 感冒(cold):P-neg 心肌炎(myocarditis) :P-neg 心电图(ECG):Pa-pos	**MGT**	感冒(cold):P-neg 心肌炎 (myocarditis) :P-neg 心电图(ECG):P-pos

Fig. 7. Two examples of clinical conversation windows and their corresponding target medical terms and status. The right part are results produced by baselines and our model. The first example comes from the CMDD dataset, while the second comes from the Chunyu dataset.

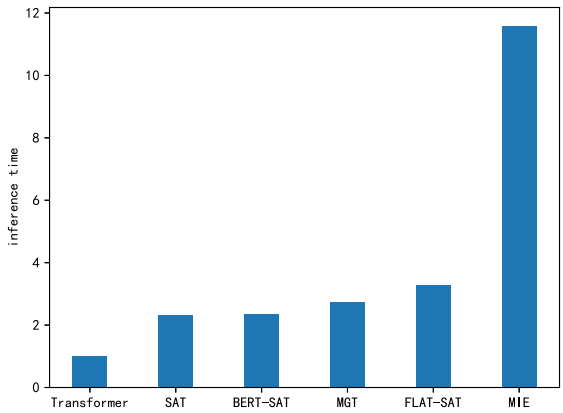
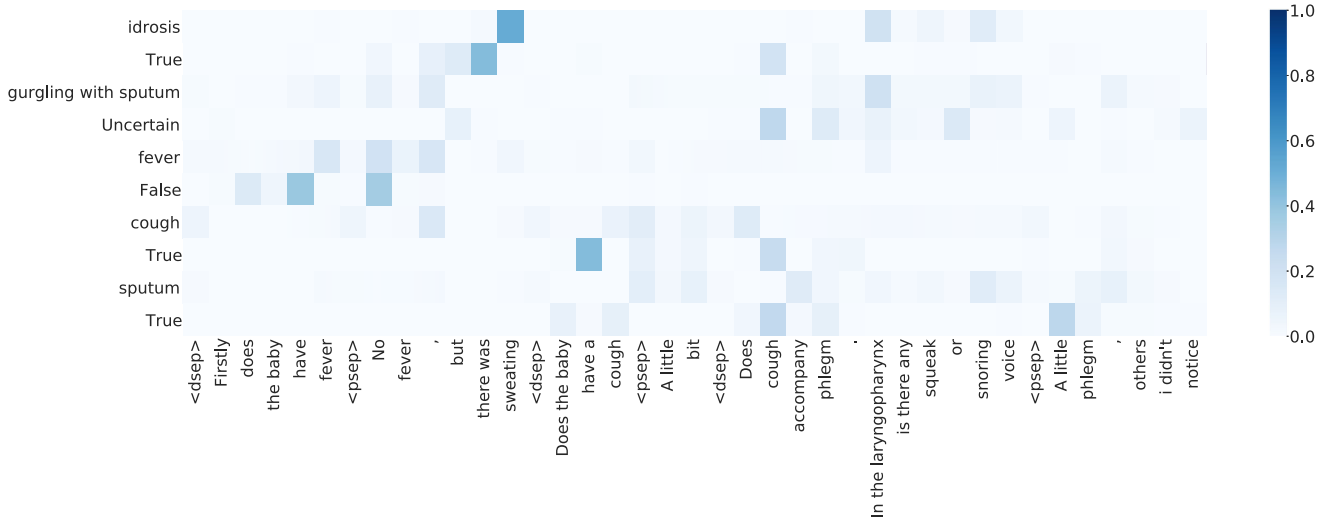
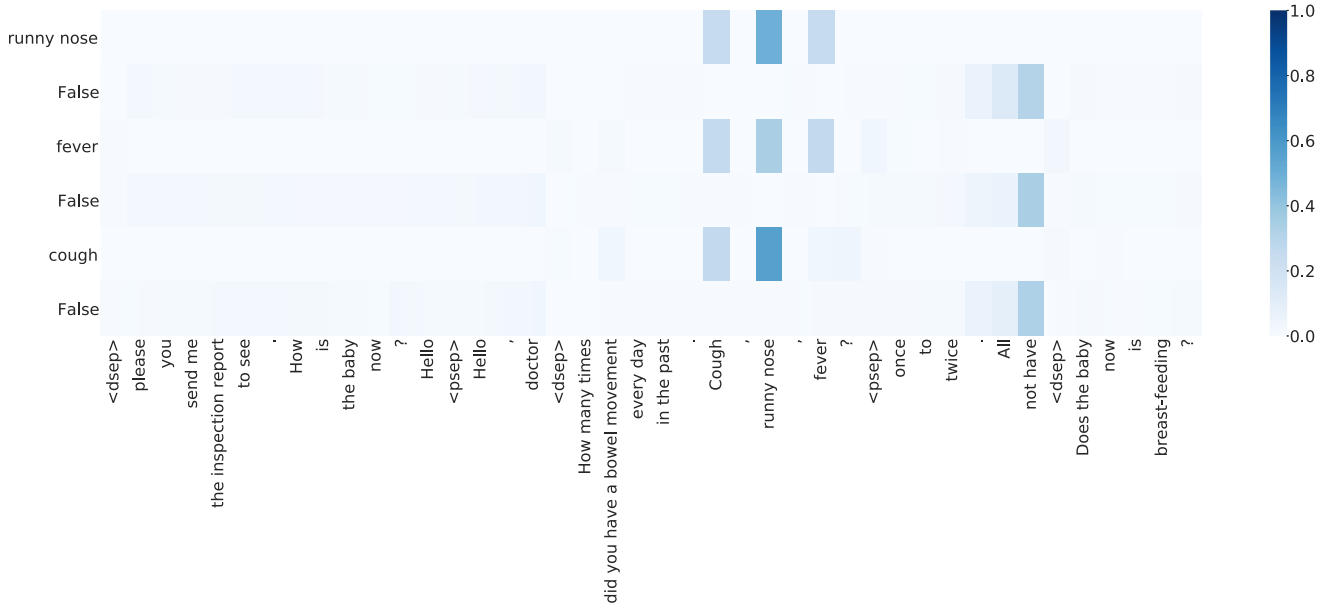


Fig. 8. Inference-time of different models (batch size is 8), compared with Transformer.



(a) Example I



(b) Example II

Fig. 9. The visualization of attention weights of the context-attention layer. Due to the original sentences are in Chinese, we try to correspond with the Chinese tokens in translation. The x-axis is the input of the dialogue text, and the y-axis is the output of our MGT model.

TABLE VII
RESULTS OF DIFFERENT HYPER-PARAMETERS ON CMDD DATASET

head_num	hidden_size	ffn_size	P	R	F
4	256	507	0.800	0.770	0.776
8	256	507	0.794	0.768	0.773
4	268	507	0.799	0.760	0.771
8	256	512	0.792	0.765	0.772
4	256	512	0.801	0.763	0.774

example I (Fig. 9(a)), we can observe that the attention values of the mentions of “In the laryngopharynx,” “snoring,” “voice” are relatively high, which illustrates that the model can capture the implicit symptom expression in the dialogue context. From example II (Fig. 9(b)), we can observe that symptom-related context is easy to capture, and our model can also correctly capture the status context. In this example, the expression of status is also obscure. The status of the three symptoms is described together in words :“All do not have”. Our model can correctly capture this information, indicating that the model can better understand the context of the dialogue.

VII. CONCLUSION

In this work, we proposed a Multi-granularity Transformer model, an end-to-end architecture, for Chinese clinical conversation understanding. We evaluated our models on two Chinese medical dialogue corpora. Extensive experiments show the

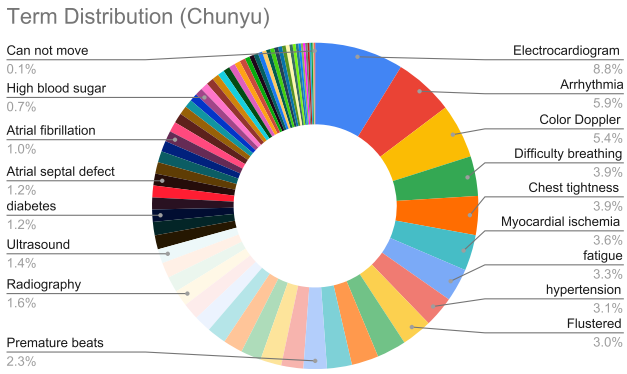


Fig. 10. The statistics of terms on the Chunyu dataset.

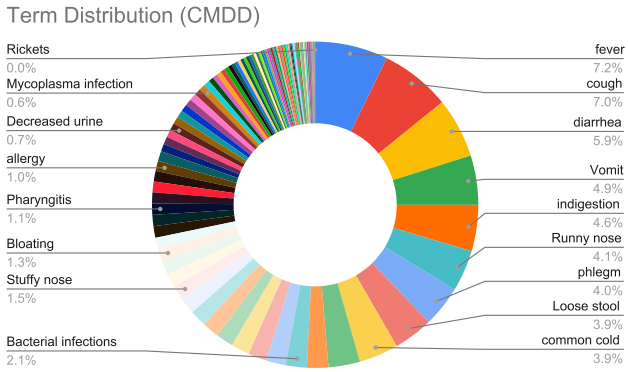


Fig. 11. The statistics of terms on the CMDD dataset.

merit of using multi-granularity information and demonstrate its power in complex dialogue scenarios, compared with the previous classification-based methods and standard generative methods. We also conduct detailed analysis to show the effect of different components. Thanks to exploring suitable dialogue interaction with our proposed mechanism RaCa, we found that the hierarchical structure is stable and surpasses other structures. All in all, we demonstrate the benefit of incorporating the multi-granularity dialogue feature in medical terms and the status generation process.

APPENDIX
ADDITIONAL STATISTICS OF TWO DATASETS

In order to better show the details of our two training corpora, we add two graphs describing their term distribution separately. The results are shown in Figs. 10 and 11.

REFERENCES

[1] A. Happe, B. Pouliquen, A. Burgun, M. Cuggia, and P. Le Beux, "Automatic concept extraction from spoken medical reports," *Int. J. Med. Informat.*, vol. 70, no. 2/3, pp. 255–263, 2003.

[2] G. Finley *et al.*, "An automated medical scribe for documenting clinical encounters," in *North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., Demonstrations*. New Orleans, Louisiana: ACL, 2018, pp. 11–15.

[3] G. Finley *et al.*, "From dictations to clinical reports using machine translation," *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA: ACL, vol. 3, pp. 121–128, 2018.

[4] P. Patel, D. Davey, V. Panchal, and P. Pathak, "Annotation of a large clinical entity corpus," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium: ACL, 2018, pp. 2033–2042.

[5] J. C. Quiroz, L. Laranjo, A. B. Kocaballi, S. Berkovsky, D. Rezazadegan, and E. Coiera, "Challenges of developing a digital scribe to reduce clinical documentation burden," *NPJ Digit. Med.*, vol. 2, no. 1, p. 114, 2019.

[6] W. Liu, J. Tang, J. Qin, L. Xu, Z. Li, and X. Liang, "Meddg: A large-scale medical consultation dataset for building medical dialogue system," in *Proc. EMNLP*, 2020, pp. 9241–9250.

[7] Z. Wei *et al.*, "Task-oriented dialogue system for automatic diagnosis," in *Proc. Assoc. Comput. Linguistics*, Melbourne, Australia: ACL, 2018, pp. 201–207.

[8] H. Kao, K. Tang, and E. Y. Chang, "Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 2305–2313.

[9] L. Xu, Q. Zhou, K. Gong, X. Liang, and L. Lin, "End-to-end knowledge-routed relational dialogue system for automatic diagnosis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7346–7353.

[10] Y. Peng, K. Tang, H. Lin, and E. Y. Chang, "REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis," in *NIP*, 2018, pp. 7333–7342.

[11] Y. Zhang *et al.*, "MIE: A medical information extractor towards medical dialogues," in *Proc. Assoc. Comput. Linguistics.* ACL, 2020, pp. 6460–6469.

[12] N. Du, K. Chen, A. Kannan, L. Tran, Y. Chen, and I. Shafran, "Extracting symptoms and their status from clinical conversations," in *Proc. Assoc. Comput. Linguistics*. Florence, Italy: ACL, 2019, pp. 915–925.

[13] N. Du, M. Wang, L. Tran, G. Lee, and I. Shafran, "Learning to infer entities, properties and their relations from clinical conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process.-Int. Joint Conf. Natural Lang. Process.*. Hong Kong, China: ACL, 2019, pp. 4979–4990.

[14] X. Lin, X. He, Q. Chen, H. Tou, Z. Wei, and T. Chen, "Enhancing dialogue symptom diagnosis with global attention and symptom graph," in *Proc. Conf. Empirical Methods Natural Lang. Process.-Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China: ACL, 2019, pp. 5033–5042.

[15] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. Assoc. Comput. Linguistics*. Berlin, Germany: ACL, 2016, pp. 1715–1725.

[16] X. Li, H. Yan, X. Qiu, and X. Huang, "FLAT: Chinese NER using flat-lattice transformer," in *Proc. Assoc. Comput. Linguistics.* ACL, 2020, pp. 6836–6842.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*. Minneapolis, Minnesota: ACL, 2019, pp. 4171–4186.

[18] J. Yan, Y. Wang, L. Xiang, Y. Zhou, and C. Zong, "A knowledge-driven generative model for multi-implication chinese medical procedure entity normalization," in *Proc. Conf. Empirical Methods Natural Lang. Process. Assoc. Comput. Linguistics*, 2020, pp. 1490–1499.

[19] S. P. Selvaraj and S. Konam, "Medication regimen extraction from medical conversations," in *Proc. Explainable AI Healthcare Med.*, 2021, pp. 195–209.

[20] T. H. Payne, W. D. Alonso, J. A. Markiel, K. Lybarger, and A. A. White, "Using voice to create hospital progress notes: Description of a mobile application and supporting system integrated with a commercial electronic health record," *J. Biomed. Informat.*, vol. 77, pp. 91–96, 2018.

[21] M. X. Chen *et al.*, "The best of both worlds: Combining recent advances in neural machine translation," in *Proc. Assoc. Comput. Linguistics*. Melbourne, Australia: ACL, 2018, pp. 76–86.

[22] R. Gangi Reddy, D. Contractor, D. Raghu, and S. Joshi, "Multi-level memory for task oriented dialogs," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, Minnesota: ACL, 2019, pp. 3744–3754.

[23] B. Peng *et al.*, "Few-shot natural language generation for task-oriented dialog," in *Findings Proc. Assoc. Comput. Linguistics, Proc. Conf. Empirical Methods Natural Lang. Process.*, ACL, 2020, pp. 172–182.

[24] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 20179–20191.

[25] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," in *Proc. Assoc. Comput. Linguistics*. Florence, Italy: ACL, 2019, pp. 808–819.

[26] L. Ren, K. Xie, L. Chen, and K. Yu, "Towards universal dialogue state tracking," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: ACL, 2018, pp. 2780–2786.
[27] S. Kim, S. Yang, G. Kim, and S.-W. Lee, "Efficient dialogue state tracking by selectively overwriting memory," in *Proc. Assoc. Comput. Linguistics*, ACL, 2020, pp. 567–582.
[28] H. Le, S. C. Hoi, and R. Socher, "Non-autoregressive dialog state tracking," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 1–21.
[29] C.-S. Wu, S. C. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, ACL, 2020, pp. 917–929.
[30] G. Tang, M. Müller, A. Rios, and R. Sennrich, "Why self-attention? a targeted evaluation of neural machine translation architectures," in *Proc. Conf. Empirical Methods Natural Lang. Process*., Brussels, Belgium: ACL, 2018, pp. 4263–4272.
[31] E. Dinan *et al.*, "The second conversational intelligence challenge (convai2)," in *Proc. NeuroIPS*, 2019, pp. 187–208.
[32] C. Alberti, K. Lee, and M. Collins, "A BERT baseline for the natural questions," 2019. *arXiv:1901.08634*.
[33] S. Mehri and M. Eskenazi, "Multi-granularity representations of dialog," in *Proc. Conf. Empirical Methods Natural Lang. Process.-Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China: ACL, 2019, pp. 1752–1761.
[34] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," in *Proc. Assoc. Comput. Linguistics*, Melbourne, Australia: ACL, 2018, pp. 1554–1564.
[35] R. Ma, M. Peng, Q. Zhang, Z. Wei, and X. Huang, "Simplify the usage of lexicon in chinese NER," in *Proc. Assoc. Comput. Linguistics*, ACL, 2020, pp. 5951–5960.
[36] J. Yan, Y. Wang, L. Xiang, Y. Zhou, and C. Zong, "A knowledge-driven generative model for multi-implication chinese medical procedure entity normalization," in *Proc. Conf. Empirical Methods Natural Lang. Process*., ACL, 2020, pp. 1490–1499.
[37] Y. Zhang, Y. Wang, and J. Yang, "Lattice LSTM for chinese sentence representation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1506–1519, 2020.
[38] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *Proc. Conf. Empirical Methods Natural Lang. Process*., ACL, 2020, pp. 6442–6454.
[39] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA: ACL, 2018, pp. 464–468.
[40] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 1–13.
[41] J. Ainslie *et al.*, "Etc: Encoding long and structured inputs in transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 268–284.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., IEEE Computer Society*, 2016, pp. 770–778.
[43] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *Proc. Deep Learn. Symp.*, 2016.
[44] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. Assoc. Comput. Linguistics*, Florence, Italy: ACL, 2019, pp. 2978–2988.
[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIP*, 2013, pp. 3111–3119.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–15.