A Speaker-Aware Co-Attention Framework for Medical Dialogue Information Extraction

Yuan Xia[1†], **Zhenhui Shi**[1†], **Jingbo Zhou**[1*], **Jiayu Xu**[1], **Chao Lu**[1], **Yehui Yang**[1], **Lei Wang**[1], **Haifeng Huang**[1], **Xia Zhang**[2], **Junwei Liu**[1]
[1]Baidu Inc., China. [2]Neusoft Corporation, China.
[1]{xiayuan,shizhenhui,zhoujingbo,xujiayu03,luchao,yangyehui01,
wanglei15, huanghaifeng, liujunwei}@baidu.com,
[2]zhangx@neusoft.com

Abstract

With the development of medical digitization, the extraction and structuring of Electronic Medical Records (EMRs) have become challenging but fundamental tasks. How to accurately and automatically extract structured information from medical dialogues is especially difficult because the information needs to be inferred from complex interactions between the doctor and the patient. To this end, in this paper, we propose a speaker-aware co-attention framework for medical dialogue information extraction. To better utilize the pre-trained language representation model to perceive the semantics of the utterance and the candidate item, we develop a speaker-aware dialogue encoder with multi-task learning, which considers the speaker's identity into account. To deal with complex interactions between different utterances and the correlations between utterances and candidate items, we propose a co-attention fusion network to aggregate the utterance information. We evaluate our framework on the public medical dialogue extraction datasets to demonstrate the superiority of our method, which can outperform the state-of-the-art methods by a large margin.

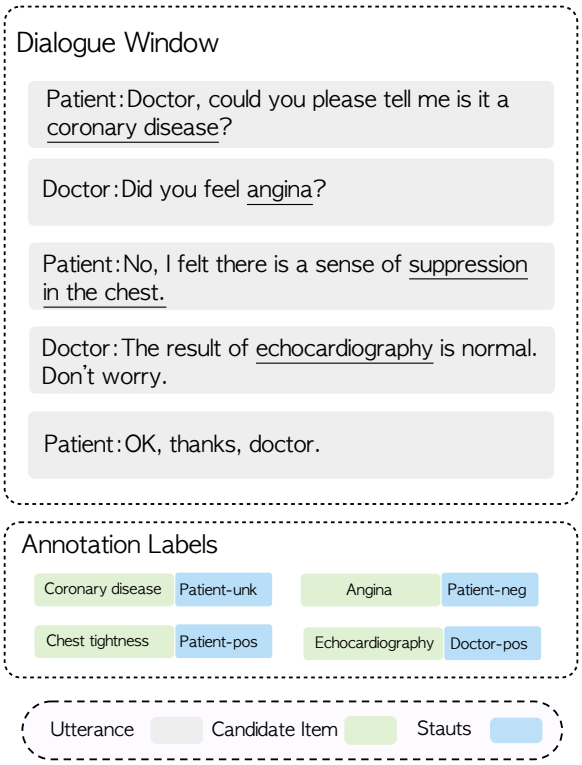

Figure 1: An example of a patient-doctor dialogue and the corresponding annotated labels.

1 Introduction

In the past decade, the collection and usage of Electronic Medical Records (EMRs) have been proved as one of the most important applications in the process of medical digitization. However, the recording and writing of the EMRs may bring a significant burden to doctors. Given the breakthrough advance of speech recognition technology, conversations between doctors and patients can be accurately recorded as text. However, such unstructured medical dialogue data cannot be easily utilized for medical research. How to automatically extract the structured information from these unstructured textual medical dialogue data is an essential step to accelerate medical digitization.

Compared with the general medical information extraction, the crucial challenge of the medical dialogue extraction is that it has to take the speaker's identity and utterance interactions into consideration. In conventional information extraction, a relation can largely be inferred by a sentence or a paragraph. However, in the medical dialogue extraction task, the candidate item and status information need to be detected and then verified by the conversations between the doctor and the patient. An example of a patient-doctor dialogue and the corresponding annotated labels is shown in Figure 1. For instance, the doctor asks the patient, *"Did*

[†]Equal contribution. *Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4777–4786
December 7-11, 2022

you feel angina?", the patient responds, *"No, I felt there is a sense of suppression in the chest."*, the ground truth labels for correct extraction are (*chest tightness: patient-positive*), (*angina: patient-negative*). If only considering the utterance of the patient or the doctor alone, we cannot make correct information extraction.

However, how to leverage the speaker's identity and utterance interactions information to facilitate medical information extraction is not well explored. Du et al. (2019) describe a novel model that extracts the symptoms mentioned in clinical conversations along with their status. The annotation of their status does not consider the speaker's identity into account. Lin et al. (2019) make symptom recognition and symptom inference in medical dialogues, and propose a global attention mechanism to capture symptom-related information. Zhang et al. (2020) develop a medical information extractor based on a simple deep matching module to take turn interaction into consideration. Thus, all existing methods fail to take the speaker into consideration, and the simple utterance combination method such as just concatenating all utterances together with flat attention cannot grasp sufficient information among utterance interactions in the medical dialogue.

To tackle the above challenge, we propose a Speaker-aware co-Attention Framework for medical dialogue Extraction (name SAFE for short). First, to better predict the status of the candidate item in the medical dialogue, we should both consider the contextual information from the dialogue and be aware of the identity of the speaker. For the annotated label (*echocardiography: doctor-positive*) in the dialogue shown in Figure 1, being aware of the identity (*patient* or *doctor*) of the speaker can help make a correct inference. Second, we propose an utterance-based co-attention graph network to perceive complex correlations between different utterances.

We summarize our contributions as follows:

- We propose a new framework (SAFE) for medical dialogue extraction, which can better utilize the pre-trained language representation model to grasp the semantics of both utterances and candidate items.
- We develop a novel speaker-aware encoder and a co-attention fusion method with multi-task learning and graph networks, which takes the speaker's identity and correlations between utterances and candidate items into consideration.
- We evaluate our framework on the public medical dialogue datasets to demonstrate the superiority of our method, which can outperform the state-of-the-art methods by a large margin.

2 Related Work

2.1 Pre-trained Language Models

Pre-trained language models, like BERT (Devlin et al., 2019), Roberta (Liu et al., 2019), XLNet (Yang et al., 2019), ERNIE (Sun et al., 2020), T5(Raffel et al., 2020), BART(Peng et al., 2021) and GPT3 (Brown et al., 2020), can achieve huge gains on many Natural Language Processing (NLP) tasks, such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. In our proposed framework, we utilize the fine-tuned BERT model as the initial encoder to obtain the representations for the utterance and the candidate item.

2.2 Medical Dialogue Extraction

Extracting information from EMR texts has attracted much research attention in both NLP and biomedical domains (Xia et al., 2021). Du et al. (2019) propose a span-attribute tagging (SAT) model and a variant of the sequence-to-sequence model to solve the symptom tagging and extraction problems. Lin et al. (2019) present a global attention mechanism, which perceives the symptom-related information from both dialogues and corpus to improve the performance of symptom recognition and symptom inference. However, the above works mainly focus on the sequential labeling and medical name entity recognition (NER), and fail to consider the complex interaction between utterances. In industrial applications, Peng et al. (2021) propose a dialogue-based information extraction system that integrates existing NLP technologies for medical insurance assessment, while their motivation is to reduce the time cost of the insurance assessment procedure.

The most similar work related to to our study is (Zhang et al., 2020), which proposes a medical information extractor (MIE) by using an LSTM (Hochreiter and Schmidhuber, 1997) model as an encoder module, and then adopting an aggregate module to take the utterance interaction into consideration. Our study is different from (Zhang et al., 2020) in the following two points. On the one

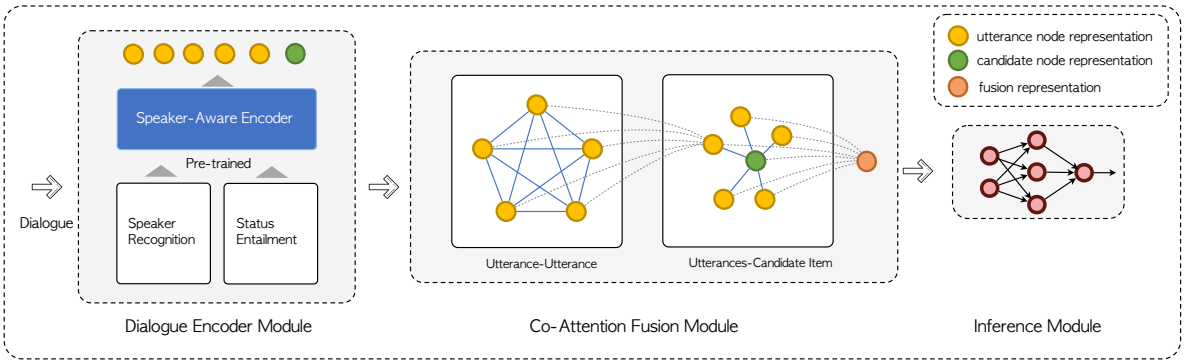

Figure 2: The Illustration of the Speaker-Aware Co-Attention Framework for Medical Dialogue Extraction (SAFE). It includes a three-stage pipeline system: a Speaker-Aware Dialogue Encoder Module (SAE), a Co-Attention Fusion Module (CAF), and an Inference Module (IM).

hand, we develop a multi-task learning method to train a speaker-aware dialogue encoder module that takes the speaker's information into consideration. On the other hand, we utilize a co-attention fusion mechanism to perceive complex interactions between different utterances and the correlation with the candidate item.

3 Preliminaries

In this section, we formally define the problem of medical dialogue extraction (MDE). For a dialogue with n tokens and m utterances, it can be defined as $D = (U_1^{r_1}, U_2^{r_2}, \cdots, U_m^{r_m})$, where $U_i^{r_i}$ is the i-th utterance in the dialogue, $r_i \in \{0, 1\}$, which indicates the speaker's identity (e.g. belongs to *patient* or *doctor*). The candidate item $I \in \mathcal{I}$ is a medical term (like symptom, disease, surgery, etc.) which can be extracted from a dialogue D. For each candidate item I, we also need to identify its status $S \in \mathcal{S}$ where S is an element from the set {*patient-negative, patient-positive, patient-unknown, doctor-positive, doctor-negative*} which indicates whether the candidate item is confirmed or denied by doctors and patients.

Finally, we define the task of medical dialogue extraction as follows: given a medical dialogue $D \in \mathcal{D}$, candidate item $I \in \mathcal{I}$ and its status $S \in \mathcal{S}$, the MDE can be formulated to predict the label $f : D \rightarrow \mathcal{Y}$ where $\mathcal{Y}$ is a matrix generated by Cartesian product of the candidate item $\mathcal{I}$ and its status $\mathcal{S}$, i.e. $\mathcal{Y} = (y_{ij}) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{S}|}$, and $y_{ij} = 1$ indicates that the medical dialogue D contains the candidate I_i with the status S_j. Note that different from the task for relation extraction (RE), the label space for the MDE is very sparse, which causes it a more challenging problem.

4 Method

We develop a three-stage pipeline system: (1) Speaker-Aware Dialogue Encoder Module (SAE), a module to turn the utterances in the medical dialogue and the candidate item into node feature representations, which also takes the speaker identity into account; (2) Co-Attention Fusion Module (CAF), a module to involve the interactions between the utterances and the correlation between utterance and candidate item into consideration; and (3) Inference Module (IM), a module to utilize the fusion representation for final dialogue information extraction. The full pipeline of our proposed medical dialogue extraction framework is illustrated in Figure 2.

4.1 Speaker-Aware Dialogue Encoder Module

An effective medical dialogue encoder should capture the semantics of the utterance and perceive the speaker's identity. In this work, we designed a multi-task learning method to pre-train our speaker-aware dialogue encoder. Our dialogue encoder is pre-trained on a *Speaker Recognition Task (SRT)* and a *Status Entailment Task (SET)*. For the *SRT* task, we design a speaker recognition task to distinguish the identity of the speaker. For the *SET* task, we leverage the pre-trained language model like BERT to train a status entailment task to perceive the semantics in the dialogue. In Figure 3, we illustrated the training process of our speaker-aware dialogue encoder module.

Speaker Recognition Task

Given an utterance in a dialogue, if the encoder itself can be aware of whether the speaker is a patient or a doctor, it will help to infer the corre-

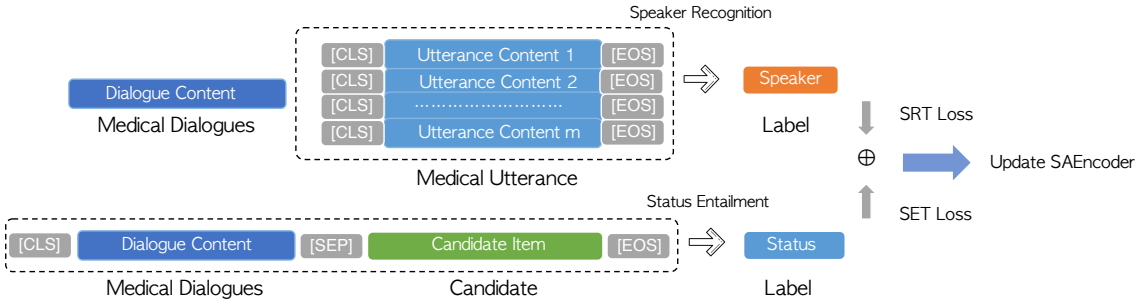

Figure 3: Illustration of the multi-task fine-tuning of Speaker-Aware Dialogue Encoder.

sponding status for the candidate item. We pre-train the BERT-base encoder with the auxiliary speaker recognition task, which is designed to distinguish whether the utterance in the medical dialogue is spoken by the patient or by the doctor. The speaker recognition task is illustrated in the upper side of Figure 3. We construct the binary training samples from the medical dialogues corpus. The utterances from the patient are labeled as 1, and the utterances from the doctor are labeled as 0. We mask the word *patient* and *doctor* at the beginning of each utterance, which can prevent the model from distinguishing the speaker only with the beginning prompt words.

First, we take the utterance U^r into the BERT-base encoder to get the utterance representation $\mathbf{U}^B$:

$$\mathbf{U}_i^B = \text{Encoder}^{(\text{BERT})}(U_i^{r_i}). \quad (1)$$

Then, we fed the utterance representation into a binary classifier, which is imposed of a dense layer and a softmax layer. The speaker recognition probability is as follows:

$$P(r_i = 1|U_i^{r_i}) = \text{softmax}(\mathbf{W}_r \mathbf{U}_i^B), \quad (2)$$

where $\mathbf{W}_r \in \mathbb{R}^{2\times d}$ denotes weight matrix, d is the number of hidden dimensions of the encoder. The loss function of the SRT for a single dialogue is as follows:

$$\begin{aligned}\mathcal{L}_{SRT} = \frac{1}{M}\sum_i & -r_i \log P(r_i = 1|U_i^{r_i}) \\ & -(1-r_i)\log P(r_i = 0|U_i^{r_i}).\end{aligned} \quad (3)$$

where M is the number of utterances in a dialogue, and r_i is the label of the speaker.

Status Entailment Task

We jointly pre-train the BERT encoder with another auxiliary status entailment task, which is designed to entail the status of the candidate item. The status entailment task is illustrated at the bottom of Figure 3. We re-formulate the medical dialogue information extraction into a status entailment task. Given a medical dialogue and the candidate item, we need to entail the status of the candidate item. The model should make an inference on the candidate's status conditioned on the dialogue and candidate item information.

First, we concatenate all the utterances in a medical dialogue D and the candidate item I together, and fed them into the BERT-base encoder to get the dialogue representation $\mathbf{D}^B$:

$$\mathbf{D}^B = \text{Encoder}^{(\text{BERT})}(D, I). \quad (4)$$

Then, we fed the dialogue representation into a multi-class (multi-status) classifier, which is also imposed of a dense layer and a softmax layer. The status entailment probability is as follows:

$$P(y|D, I) = \text{softmax}(\mathbf{W}_e \mathbf{D}^B), \quad (5)$$

where $\mathbf{W}_e \in \mathbb{R}^{C\times d}$ denotes weight matrix, d is the number of hidden dimensions of the encoder, C is the number classes of the status. The loss function for the SET is as follows:

$$\mathcal{L}_{SET} = \text{CrossEntropy}(y, P(y|D, I)). \quad (6)$$

where y is ground truth status label for candidate item in the dialogue, and $\text{CrossEntropy}(\cdot)$ is cross entropy loss function.

Joint Optimizing

The final loss function for the speaker-aware encoder $\text{Encoder}^{(\text{SA})}$ is as follows:

$$\mathcal{L}_{SAE} = \lambda \mathcal{L}_{SRT} + (1-\lambda)\mathcal{L}_{SET}. \quad (7)$$

where $\mathcal{L}_{SRT}$ and $\mathcal{L}_{SET}$ are the losses for speaker recognition task and status entailment task, respectively, λ is the hyper-parameter to control the weight of each task.

4.2 Co-Attention Fusion Module

Given the medical dialogue, we employ our pre-trained speaker-aware encoder $\text{Encoder}^{(\text{SA})}$ as our utterance encoder by extracting the final hidden state of the `[CLS]` token as the representation, where `[CLS]` is the special classification embedding in our pre-trained model. In order to involve the correlation between the utterance and the candidate item, given m utterances $(U_1^{r_1}, U_2^{r_2}, \cdots, U_m^{r_m})$ in a dialogue and a candidate item I, we feed each utterance-candidate item pair $(U_i^{r_i}, I)$ into our speaker-aware encoder to obtain the utterance representation $\mathbf{U}_i$. We also feed the candidate item I into the speaker-aware encoder alone to obtain the candidate item representation $\mathbf{I}$:

$$\begin{aligned} \mathbf{U}_i &= \text{Encoder}^{(\text{SA})}(U_i^{r_i}, I), \\ \mathbf{I} &= \text{Encoder}^{(\text{SA})}(I), \end{aligned} \tag{8}$$

To better capture complex interactions between utterances, we use a co-attention fusion mechanism to aggregate the utterance information. We treat each utterance as a node and define other utterances in the same sliding window as its neighbors. Then we calculate the attention coefficient between a node i and its neighbor j ($j \in \mathcal{N}_i$).

$$c_{ij} = \mathbf{W}_1^{u\to u}(\text{ReLU}(\mathbf{W}_0^{u\to u}(\text{concat}(\mathbf{U}_i, \mathbf{U}_j)))), \tag{9}$$

where $j \in \mathcal{N}_i$ is the in-window neighbors of the node i, $\mathbf{W}_1^{u\to u} \in \mathbb{R}^{1\times w}$ and $\mathbf{W}_0^{u\to u} \in \mathbb{R}^{w\times 2d}$ are weight matrices, and $\text{concat}(\cdot,\cdot)$ is concatenation operation. d is the number of dimensions of the utterance feature representation, w is the number of dimensions of the intermediate hidden state.

We use a softmax function to normalize the utterance-utterance co-attention coefficients ϕ,

$$\phi_{ij} = \text{softmax}(c_{ij}) = \frac{\exp(c_{ij})}{\sum_{k\in\mathcal{N}_i}\exp(c_{ik})}. \tag{10}$$

Then, given the utterance-utterance co-attention matrix ϕ_{ij}, inspired by (Kipf and Welling, 2017; Veličković et al., 2018; Zhou et al., 2019), we employ a simple GCN layer for information fusion.

$$\widetilde{\mathbf{U}}_i^{(l)} = \sigma\Big(\sum_{j=1}^{n}\phi_{ij}\mathbf{W}_\phi^{(l)}\widetilde{\mathbf{U}}_i^{(l-1)}\Big), \tag{11}$$

where $\widetilde{\mathbf{U}}_i^{(0)}$ is initialized with $\mathbf{U}_i$, $\mathbf{W}_\phi^{(l)} \in \mathbb{R}^{d\times d}$, l is the number of layers for propagation.

We also explicitly involve the correlation between the utterance $\widetilde{\mathbf{U}}_i^{(l)}$ and the candidate item $\mathbf{I}$ by another co-attention layer:

$$p_i = \mathbf{W}_1^{u\to c}(\text{ReLU}(\mathbf{W}_0^{u\to c}(\text{concat}(\widetilde{\mathbf{U}}_i^{(l)}, \mathbf{I})))), \tag{12}$$

where $\mathbf{W}_1^{u\to c} \in \mathbb{R}^{1\times w}$ and $\mathbf{W}_0^{u\to c} \in \mathbb{R}^{w\times 2d}$ are weight matrices.

Similarly, we adopt a softmax function to normalize the utterance-candiate item co-attention coefficients ψ,

$$\psi_i = \text{softmax}(p_j) = \frac{\exp(p_i)}{\sum_{k=1}^{N}\exp(p_k)}, \tag{13}$$

Finally, the normalized co-attention coefficients are used to compute a linear combination of utterance features of neighbors for final information extraction:

$$\mathbf{T}_F = \text{CoAttn}(D, I) = \sum_{k=1}^{N}\psi_k\widetilde{\mathbf{U}}_k^{(l)}. \tag{14}$$

4.3 Inference Module

The output representation $\mathbf{T}_F$ of the Co-Attention Fusion module (CAF) is then fed into the final inference module to extract the medical information from the dialogue.

$$\tilde{y}_c = \text{softmax}(\mathbf{W}_o\mathbf{T}_F^{(c)} + \mathbf{b}_o), \tag{15}$$

where $\mathbf{T}_F^{(c)}$ is the c-th index of the candidate item, $\mathbf{W}_o \in \mathbb{R}^{C\times d}$ and $\mathbf{b}_o \in \mathbb{R}^{C\times 1}$ are weight matrix and bias, respectively. $\tilde{y}_c$ is the predicted probability of the candidate item's status, y_c is the ground-truth label.

The final loss function is as follows:

$$\mathcal{L} = \frac{1}{NC}\sum_i\sum_c y_c^{(i)}\log\tilde{y}_c^{(i)}. \tag{16}$$

where N is number of dialogues in the training corpus, C is the number of classes for candidate item status.

5 Experiments

5.1 Datasets

To verify the effectiveness of our SAFE framework, we conduct extensive experimental evaluations on the Medical Information Extraction **MIE** dataset[1]

[1]Data are available at https://github.com/nlpir2020/MIE-ACL-2020

(Zhang et al., 2020). The dataset involves doctor-patient dialogues collected from a Chinese medical consultation website [2]. The MIE dataset is representative for medical dialogue task from EMR. On the one hand, the dialogues from the MIE dataset are collected from real doctor-patient conversations, it can reflect the data characteristics from EMRs. On the other hand, for industrial applications, the problem of extracting and structuring of EMRs raised by the MIE dataset has become a fundamental task in downstream medical applications, such as text-based dialogue systems or cascaded with ASR (Automatic Speech Recognition) systems.

In the MIE dataset, the dialogues are already in text format. As the dialogues turn to be too long, the medical dialogues are processed into pieces using a sliding window. A window consists of multiple consecutive turns of a dialogue. The sliding window size is set to 5, because this size allows the included dialogue turns contain proper amount of information. For windows with less than 5 utterances, the dataset pads them at the beginning with empty strings. Then, it uses a window-to-information annotation method, and annotates the candidate item and its status in each window in the dialogue. Annotators of the MIE dataset are guided by two physicians to ensure the correctness and the cohen's kappa coefficient of the labeled data is 0.91. It defines four categories (i.e. *symptom*, *surgery*, *test*, and *other information*) and 71 candidate items which are frequent items in doctor-patient dialogues and are fixed in the MIE dataset. The candidate item has five statuses (i.e. *patient-pos*, *patient-neg*, *doctor-pos*, *doctor-neg*, *patient-unknown*). In total, the corpus has 1,120 dialogues and 18,212 windows. For the dialogue-level, the dataset is split into three parts: training, validation and testing, and the sizes are 800, 160, and 160, respectively; for the window-level, the corresponding sizes are 12,932, 2,587, and 4,254, respectively. The detailed annotation statistics of the MIE dataset are shown in Table 1.

	Train	Dev	Test
# Window-level	12,932	2,587	4,254
Avg. words of windows	110.8	113.3	109.7
Avg. annotations of windows	2.5	2.7	2.4
# Dialogue-level	800	160	160
Avg. words of dialogues	404.4	434.7	401.3
Avg. annotations of dialogues	6.5	7.2	6.4

Table 1: The detailed annotation statistics of the MIE dataset.

5.2 Evaluation Metrics

For the MIE dataest, we evaluate the extracted medical dialogue information results with *Precision*, *Recall* and *F1-Score*. In accordance with the evaluation metrics described in the (Zhang et al., 2020), a correct result should both correctly predict the candidate item and its status. The results are evaluated in window-level and dialogue-level as follows:

- **Window-level.** The evaluation is calculated with each segmented window, and report the micro-average of all the test windows.
- **Dialogue-level.** First, we merge the results of windows belonging to the same conversation. For mutually exclusive status, we update the previous status with the latest status. Then, we evaluate the results of each dialogue and report the micro-average of all test dialogues.

5.3 Experiment Settings

Task Training Settings

For the speaker recognition task, the label of the speaker in each utterance is generated by the beginning prompt words (e.g. *patient:* or *doctor:*). In the training stage, we mask the beginning prompt words to prevent the leakage of labels. For the status entailment task, in addition to the origin status labels (e.g. *patient-pos*), we add the *None* status label as the *negative* label. Because the candidate item is not provided in the inference stage, thus we have to traverse the candidate item space to make a prediction. For a given dialogue and the provided candidate item-status pair information, suppose there are B candidate items labels presented in a dialogue, we randomly select $N_s \times B$ items which are not presented in the ground-truth candidate items and label them with the *None* status. In our experiments, we set N_s as 2. In the inference stage, we make prediction on the whole candidate item space. Only the candidate item with non-*None* status is left for final evaluation.

Hyperparameter Settings

For the speaker-aware dialogue encoder module, we use a BERT-base network structure to initialize the base dialogue encoder. The BERT-base (110M) namodel has 12 layers, the number of hidden state dimensions is set to 768, and the number of heads

[2] https://www.chunyuyisheng.com

Method	Window-Level			Dialogue Level		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
LSTM-Classifier	53.13	49.46	50.69	61.34	52.65	56.08
MIE-Single (Zhang et al., 2020)	69.40	64.47	65.18	75.37	63.17	67.27
MIE-Multi (Zhang et al., 2020)	70.24	64.96	66.40	**76.83**	64.07	69.28
MIE-Multi (BERT)	71.45	71.17	71.31	71.01	74.46	72.69
SAFE (Ours)	**72.59***	**73.86***	**73.22***	73.20	**78.71***	**75.86***

Table 2: Performance comparisons with different baseline models on the MIE dataset. Significant test over the best baseline results are marked with * (pair-wised t-test , $p < 0.01$).

is set to 12. We use the Adam optimizer (Kingma and Ba, 2015) with a batch size of 32 for 20 epochs. The learning rate α_s for SAE pre-training is set to 2e-5. The warmup proportion is set to 0.1. The maximum sequence length is 512. The λ for controlling the task weight is set to 0.5 with grid search strategy. For the co-attention fusion module, the number of hidden dimensions of the dense layer is set to 64, and the number of layers for utterance propagation is set to 2. The final inference module is trained to minimize the cross-entropy loss on the predicted label using the Adam optimizer with a batch size of 128 for 15 epochs, and the initial learning rate α_c for co-attention fusion method is set to 1e-3. The models are trained on the NVIDIA Tesla V100 32GB GPU with 4 hours.

5.4 Model Comparisons

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

- **LSTM-Classifer** The model only uses the LSTM encoder to get the representation of the concatenation of each utterance and uses a self-attention layer and an MLP layer to make predictions.
- **MIE-Single (Zhang et al., 2020)** The model uses the LSTM model as the encoder module, and only consider the interaction within a single utterance.
- **MIE-Multi (Zhang et al., 2020)** The model uses the LSTM model as an encoder module and proposes a simple aggregate module to take the utterance interaction into consideration.
- **MIE-Multi (BERT)** The model architecture is the same with the MIE-Multi, except that we replace the original LSTM encoder with the BERT encoder.

Method	Precision	Recall	F1-Score
SAFE	73.20	78.71	**75.86**
w/o (SAE)	68.71	78.51	73.29
w/o (CAF)	69.46	74.55	71.91

Table 3: The ablation study on the MIE dataset with dialogue-level metrics.

CAF Layers	Precision	Recall	F1-Score
1	71.91	79.10	75.34
2	73.20	78.71	**75.86**
3	71.50	76.53	73.93

Table 4: Performance of different number of co-attention layers in the CAF module on the MIE dataset with dialogue-level metrics.

- **SAFE (Ours)** Our speaker-aware co-attention framework takes the speaker's identity and the correlations between utterances and candidate items into consideration.

5.5 Main Results

In accordance with the evaluation metrics introduced by Zhang et al. (2020), we report both window-level and dialogue-level results. Table 2 shows the performance comparisons with different methods on the MIE dataset. We observe that the LSTM-Classifier performs the worst, under the dialogue-level metrics. The LSTM-Classifier only has a precision of 61.34 and a recall of 52.65, because it fails to consider interactions between each utterance. The performance of the MIE-Multi is better than the MIE-Single, as the latter model takes the turn interactions into account. The MIE-Multi achieves better performance at a precision of 76.83 and a recall of 64.07 under the dialogue-level metrics. The MIE-Multi is a state-of-art framework for medical dialogue extraction. However, without taking the speaker's identity into consideration, the MIE-Multi cannot tackle complex interactions between utterances and candidate items, it perform

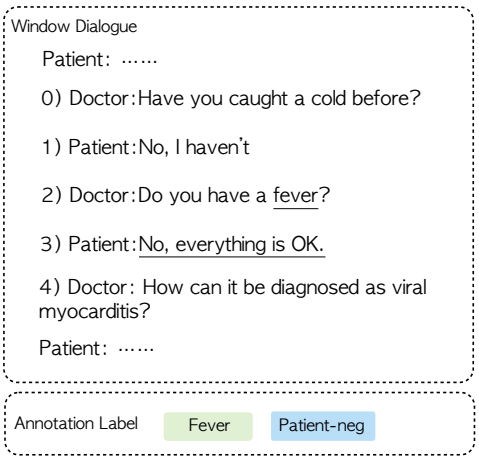

Figure 4: An case study on a patient-doctor dialogue in the test set. The corresponding utterance-utterance co-attention matrix is shown in Figure 5.

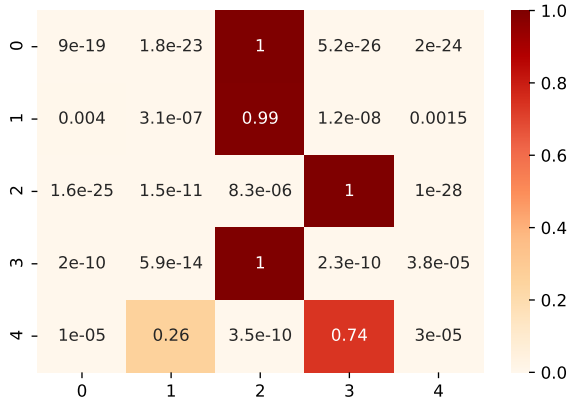

Figure 5: The co-attention matrix for the case in Figure 4. The attention map indicates the utterance-utterance interaction between different speakers.

less effectively compared to our SAFE framework.

For a more fair comparison, to eliminate the performance boost brought by pre-trained language models like BERT, we re-implement the MIE-Multi with a BERT-based structure, the MIE-Multi (BERT) gets an F1-Score of 71.31 under window-level metrics and an F1-Score of 72.69 under the dialogue-level metrics, which is better than the original MIE-Multi, while still getting worse results compare to our method. Our SAFE framework achieves the state-of-the-art F1-Score of 75.86 which demonstrates the superiority of our method by a large margin.

5.6 Ablation Study

We conduct ablation studies on the MIE dataset to analyze the contribution of each component of our proposed SAFE model. The main results are shown in Table 3.

Effectiveness of Speaker-Aware Encoder

First, we evaluate the effect of the speaker-aware encoder module. The removal of the SAE module causes the overall performance of the F1-score to decline from 75.86 to 73.29 under the dialogue-level metrics, which suggests that taking the speaker's information into account can help improve the dialogue extraction performance. Additionally, to quantitatively demonstrate that the SAE module can identify the speaker better, we calculate the speaker misidentification error rate in the test set, which indicates how many bad cases are owing to the error of speaker identity (*e.g., pred: doctor-pos, label: patient-pos*). The speaker misidentification error rate is decreased from 5.0% to 4.1% compared to the method without the SAF module.

Effectiveness of Co-Attention Fusion Module

Second, we evaluate the effect of the co-attention fusion module. Removing the CAF module reduces the overall performance of the F1-score by 5.49% (from 75.86 to 71.91) under the dialogue-level metrics, which proves that adopting the co-attention graph network to capture the complex interactions between the utterances is significant for medical dialogue extraction. We also analyze the effect of the different number of co-attention layers on the performance of medical dialogue extraction. The results are shown in Table 4. Note that when the co-attention layer is equal to 1, the CAF is equivalent to the flat attention over utterances. We can discover from the table that the model with two co-attention layers achieves the best result, which indicates that the proper propagation of each utterance can help to perceive complex interactions in the medical dialogue.

5.7 Case Study

In previous sections, we provide a quantitative analysis of the experiment results. In this section, to help better understand that our SAFE framework can better capture utterance interactions in the dialogue, we provide a case study from the test set.

Figure 4 shows a case study on a patient-doctor dialogue[3] in the test set. To illustrate how our co-attention fusion module can capture interactions between each speaker and the correlation with the candidate item, we visualize the utterance-utterance

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

interaction with an attention map. From the Figure 5, we can find that the third column (*Doctor: Do you have a fever?*) and the fourth column (*Patient: No, everything is OK.*) of the matrix have dominantly higher values, because these two utterances are important for the model to extract the annotated label *(fever: patient-negative)*. We can also discover that the co-attention coefficients (i.e. $\phi_{2,3}$ and $\phi_{3,2}$) of these two utterances are also very high, because we need to consider the interactions between these two utterances to infer the ground-truth status as *patient-negative*.

5.8 Discussion

Here we would like to give a brief discussion about how the proposed system connects with clinical practice. For text-based systems, the structured information from the text-based dialogues can be extracted to form the medical knowledge graph, which would benefit primary doctors. The structured information from medical dialogues can also bring benefits for many clinical applications, such as automatic diagnosis systems (Liu et al., 2018; Xu et al., 2019; Xia et al., 2020) and clinical decision support systems to assist doctors. For ASR systems, it is also important to utilize the speaker identity recognition in the system to facilitate medical information extraction after speech recognition.

6 Conclusion

In this paper, we propose a speaker-aware co-attention framework for medical dialogue information extraction. We design a speaker-aware dialogue encoder module, which considers the speaker's identity into account and can better utilize the pre-trained language model to capture the semantics of the utterance and the candidate item. Moreover, we propose a co-attention fusion network to aggregate the utterance information, which tackles complex interactions between different utterances and the correlation between utterances and candidate items. The experiment results demonstrate the effectiveness of the proposed framework.

7 Limitations

While perceiving the speaker's identity and complex utterance interactions is essential for medical dialogue information extraction, the limitation of our work is that we do not explicitly involve the prior medical knowledge such as the existing medical knowledge graph (MKG) to further improve the overall performance with less annotated labels. To deal with the limitation, in the future, we should leverage the medical entity relations in the medical knowledge graph, and introduce the medical knowledge enhanced pre-train language model into our work to further improve the results of medical dialogue information extraction.

8 Ethical Considerations

It should be mentioned that the doctor-patient dialogues in the MIE dataset are collected from the openly accessible online health forum Chunyu-Doctor whose owners make such information visible to the public. All the patients' information has been anonymized. Apart from the personal information de-identified by the Chunyu-Doctor forum officially, we manually reviewed the collected data to prevent privacy leaks. We ensure there is no identifiable or offensive information in the experimental dataset.

The model and framework proposed in this paper are for research purposes only and intended to facilitate studies of using NLP methods to better extract the structure information from medical dialogues, which can alleviate the doctor's burdens for recording EMRs and accelerate the development of medical digitization.

Acknowledgement

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

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, pages 4171–4186.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 915–925, Florence, Italy. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042, Hong Kong, China. Association for Computational Linguistics.

Qianlong Liu, Zhongyu Wei, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *ACL*, volume 2, pages 201–207.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shuang Peng, Mengdi Zhou, Minghui Yang, Haitao Mi, Shaosheng Cao, Zujie Wen, Teng Xu, Hongbin Wang, and Lei Liu. 2021. A dialogue-based information extraction system for medical insurance assessment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 654–663, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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*, pages 8968–8975.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yuan Xia, Chunyu Wang, Zhenhui Shi, Jingbo Zhou, Chao Lu, Haifeng Huang, and Hui Xiong. 2021. Medical entity relation verification with large-scale machine reading comprehension. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3765–3774.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1062–1069.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *AAAI*.

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*, pages 5754–5764.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. MIE: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469, Online. Association for Computational Linguistics.

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*, pages 892–901.