Matching Questions and Answers in Dialogues from Online Forums

Qi Jia [1] and **Mengxue Zhang** [2] and **Shengyao Zhang** [3] and **Kenny Q. Zhu** [4]

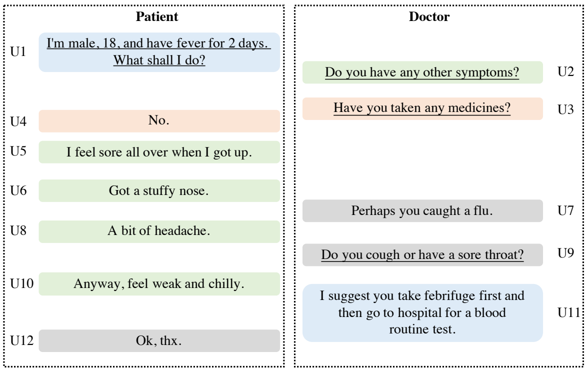
Abstract. Matching question-answer relations between two turns in conversations is not only the first step in analyzing dialogue structures, but also valuable for training dialogue systems. This paper presents a QA matching model considering both distance information and dialogue history by two simultaneous attention mechanisms called mutual attention. Given scores computed by the trained model between each non-question turn with its candidate questions, a greedy matching strategy is used for final predictions. Because existing dialogue dataset such as the Ubuntu dataset are not suitable for the QA matching task, we further create a dataset with 1,000 labeled dialogues and demonstrated that our proposed model outperforms the state-of-the-art and other strong baselines, particularly for matching long-distance QA pairs.

1 Introduction

Question motivated dialogues are very common in daily life and they are rich sources for question-answer (QA) pairs. For example, in an online forum for health consultations, both the doctor and the patient tend to ask and answer questions to narrow down the information gap to reach the final diagnosis or recommendations. Matching QA pairs can help tracking the final answer from the doctor to the orginal patient question and is valuable for the medical domain.

QA matching is an important part of analyzing discourse structures for dialogue comprehension. Asher et. al [2] shows that in online dialogues where participants are prompted to communicate with others to achieve their goals, 24.1% of the relations between elementary discourse units are QA pairs. Questions and answers are the main components of dialogue acts [30], which provide key features for dialogue summarization and decision detection [15]. Besides, figuring out the QA relations between these utterances can provide question answering models [19, 32, 9] with high-quality QA pairs and contribute to the exploration of proactive questioning [37].

However, many challenges exist. While it's relatively easy to distinguish between questions and non-questions [5], the non-questions may contain not only valid answers, but also chit-chats and other informative statements. It is also a common phenomenon that a long and complete answer is broken up into several turns such as $\{U5,U6,U8,U10\}$ in Figure 1. Due to factors such as the network delay and differences in typing speed, the dialogue sequences are always mix-matched. Moreover,"personalized" orthography, ellipses, abbreviations, and missing punctuations are all difficulties for QA matching.



U1 I'm male, 18, and have fever for 2 days. What shall I do?
U2 Do you have any other symptoms?
U3 Have you taken any medicines?
U4 No.
U5 I feel sore all over when I got up.
U6 Got a stuffy nose.
U7 Perhaps you caught a flu.
U8 A bit of headache.
U9 Do you cough or have a sore throat?
U10 Anyway, feel weak and chilly.
U11 I suggest you take febrifuge first and then go to hospital for a blood routine test.
U12 Ok, thx.

Figure 1. Questions and answers matching in dialogues from an online health forum. The identified pairs are painted in the same color and questions are underlined.

In this work, we focus on the task of matching questions and answers in two-party multi-turn dialogues. We found that the distance between the question and its answers isn't only caused by the mix-matching and fragmentation mentioned above, but by the very nature of the question. Some questions can be answered directly based on personal knowledge, such as U3, while others can not. For instance, when a patient asks questions such as "what's wrong with me" or "what should I do" just like U1, the doctor often has to ask follow-up questions $\{U2,U3\}$ to seek additional information in order to give the final diagnosis or recommendation (U11). This often has to be done with several rounds of communication. We call this kind of QA pairs *incremental QA*. Such QA pairs often form the main idea of a dialogues or sub-dialogue, critical for dialogue comprehension. The answers are inherently far from the question for incremental QA pairs (distance ≥ 3 [6]), aggravating the difficulty of matching such pairs.

Roughly, we can categorize QA pairs according to the distance between them. When distance ≤ 3, we call it short-distance QA pairs (SQA); otherwise, long-distance QA pairs (LQA). It is obvious that matching LQAs is more difficult than SQAs. We assume that a two-party multi-turn dialogue contains two types of turns, questions (Q) and non-questions (NQ), which are labeled in advance [7]. Our task is

[1] Shanghai Jiao Tong University, China , email: Jia_qi@sjtu.edu.cn
[2] Shanghai Jiao Tong University, China , email: mengxue.zhang96@gmail.com
[3] Shanghai Jiao Tong University, China , email: sophie_zhang@sjtu.edu.cn
[4] Shanghai Jiao Tong University, China , email: kzhu@cs.sjtu.edu.cn

[5] This can be done with a simple neural-based classifier with high accuracy.

[6] There is at least one follow-up question and one corresponding answer between the question and answer of an incremental QA pair. So the distance for such QA pairs is larger or equal to 3.

[7] We implemented a simple LSTM-based Q/NQ classifier with accuracy equaling 96.10% and F1-macro equaling 95.07% on this dataset which will be released to the research community. Question detection is not the focus of this paper.

to identify all answers from the set of NQs to a given Q.

Previous methods on the task [10, 12, 20] suffer from a major weaknesses: while classifying a pair of turns, they ignore the context of the turns in the dialogue. Meanwhile, their pre-defined features such as question words and answer words, are already implied by the Q and NQ labels in our definition and hence are not suitable for our task. He et al. [18] improves the above methods with a recurrent pointer network (RPN) model that takes the whole dialogue as an input. Their model was evaluated on a close-source customer service dialogue dataset. Although their model makes use of the context, they treat every utterance in the context equally with RNN-based networks which fails to capture the influences between turns especially with long distance. Besides, it encodes the distance information implicitly which downplays the effect of distance between the utterances. According to our experiments, none of the above approaches perform well on LQA pairs.

In this paper, we bring the dialogue context into the above simple models. For a given pair of Q and NQ to be matched, the context is defined as *history*, refering to the utterances between the Q and NQ. The critical part of our model is two simultaneous attention mechanisms that combine the history in a mutual way. Existing dialogue datasets, such as the Ubuntu dataset, do not contain the QA matching labels. More over, they are either not two-party dialogues, or do not contain long distance QA pairs. Therefore, we developed a new dataset based on a Chinese online medical forum. We conducted experiments on this and the part of the labeled Ubuntu dataset.

Our main contributions are as follows:

- We aggregate dialogue history and distance information into a new deep neural QA matching model. We show that distance is an important feature when encoded explicitly, and that utterances between Q and A can be effectively captured by a mutual attention mechanism (Section 3).
- Since there is no open source dialogue datasets designed for QA matching task, especially for long-distance QAs, we construct a reasonably sized dataset and will release it to the research community (Section 4).
- The experimental results show that our proposed method outperform other strong baselines, especially on LQAs. The techniques developed here are generic and can be applied to other types of online dialogues (Section 6).

2 Problem Definition

Our work aims at identifying the response turns to a question turn in a multi-turn, two-party dialogue. Given a dialogue sequence with T turns:

$$[(R_1, L_1, U_1), (R_2, L_2, U_2), ..., (R_T, L_T, U_T)]$$

where R denotes the role, identifying which party utters the turn. $L \in \{Q, NQ\}$, and U is a sequence of words that form an utterance.

Our job is to match each (Q, U_i) with corresponding (NQ, U_j), where:

$$\begin{aligned} j > i \quad 1 \leq i, j \leq T \\ R_i \neq R_j \end{aligned} \tag{1}$$

The *distance* of a Q-NQ pair (U_i, U_j) is $j-i$. We define the *history* of such a Q-NQ pair as the turns $\{U_{i+1}, U_{i+2}..., U_{j-1}\}$ which are located between the Q and NQ. The intuition will be explained in Section 3.2.

Recent work by He et al. [18] considers a slightly different QA alignment problem where one answer can be matched with multiple questions. However, in this paper, we assume that if a question is asked repeatedly, the answer should be matched to the nearest question and all earlier ones should be disregarded. By our definition, a Q can match nothing (U9) or several NQs (U2). From the viewpoint of a NQ, it is either matched, or not matched, with a question (such as U7). When a NQ is matched to a question, it is considered as an *answer* (A). Otherwise, it's considered as others(O).

3 Approach

We propose an attention-based neural network model for QA matching in multi-turn dialogues between two parties. Given a Q-NQ pair with its distance and history, the model consists of four components (shown in Figure 2): *Sentence Encoder* transforms the natural language turns into sentence-level embeddings. *Mutual Attention* combines history turns based on two simultaneous attention mechanisms in an interleaving way. *Match-LSTM* is used to compare the processed sentence pair word by word. *Prediction Layer* incorporates the distance information and calculates the alignment probability. After calculating the probability for all Q-NQ pairs in a dialogue, a *greedy matching algorithm* is employed to match each NQ to zero or one Q.

3.1 Sentence Encoder

Given an input turn as a sequence of words in natural language, we generate a neural representation using an LSTM [16]. The input layer consists of pretrained word embeddings of the words which are fed into a single hidden layer. The output of all the hidden states or the last hidden state can both be regarded as the sentence-level embedding for this turn.

With the sentence encoder, we can get the neural representations for a Q-NQ pair and its history as follows:

$$\begin{aligned} Q &= \{h_i^q\}_{i=1}^N \\ NQ &= \{h_i^p\}_{i=1}^M \\ H_{RQ} &= \{d_t^q\}_{t=1}^A \\ H_{RNQ} &= \{d_t^p\}_{t=1}^B \end{aligned} \tag{2}$$

where the number of words in Q and NQ are N and M respectively. h^i represents the hidden state of each word. H_{RQ} and H_{RNQ} represents the history turns with the same role label as Q and NQ respectively. A and B are the number of turns in each H_{RQ} and H_{RNQ} respectively, and d_t is the last hidden state of each turn. Superscripts p and q are used to distinguish Q and NQ. Here, we divide the history turns into two parts to support for the idea of mutual attentions between Q and NQ in the following subsection.

The intuition for using different granularity of sentence embeddings is that we hope to keep more information for more important turns. Therefore, in order to calculate the matching score between a Q-NQ pair, we preserve all the hidden states for Q and NQ. The last hidden state of each turn in history is used to provide auxiliary information.

3.2 Mutual Attention to the History

To improve the prediction for each Q-NQ pair, naturally we take advantage of the dialogue context, especially the turns between Q and

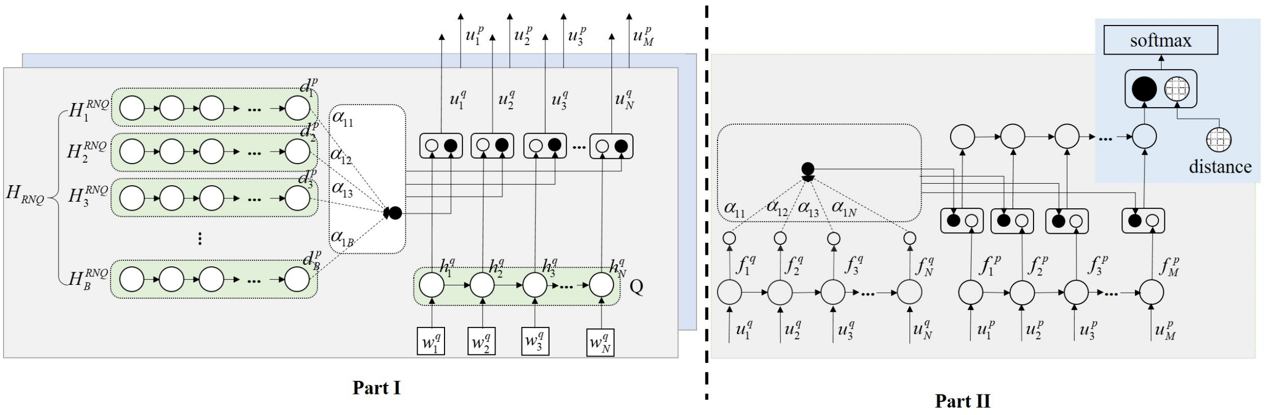


Figure 2. The architecture of the proposed match-LSTM based model with parallel attention mechanisms. **Part I:** Sentence Encoder and Mutual Attention. Front plate (gray) shows the encoding of Q, while the back plate (blue) includes the encoding of a NQ. The green part is Sentence Encoder and the rest is Mutual Attention. **Part II:** Match-LSTM and Prediction Layer. The gray part is Match-LSTM and the blue part is Prediction Layer.

NQ. This idea comes from two considerations: i) if Q has been partially answered by another NQ between the current pair of Q and NQ, then we should further explore whether the current NQ is a supplementary answer; ii) if there exists another Q which is closer to the NQ in both distance and semantics, the probability of matching the current Q-NQ pair should reduce. In a word, if the model can capture these intuitions, it's more likely to match the LQAs such as {U2,U10}.

Besides, it should be noted that the question and matching answer should definitely be uttered by different parties. In other words, the QA relations in a dialogue is focusing on the process of narrowing down the information gap between two parties, where the information interaction between parties is critical. So, we further divide the history into two parts: H_{RQ} and H_{RNQ} by different role labels as mentioned above. Q is expected to interact with H_{RNQ} while NQ is expected to interact with H_{RQ}.

Borrowing the idea from Wang et al. [35], we use two attention mechanisms to incorporate the history information into Q and NQ individually in a unified manner. For example, when dealing with the Q-NQ pair $\{U2, U10\}$, $H_{RQ} = \{U3, U7, U9\}$ and $H_{RNQ} = \{U4, U5, U6, U8\}$. The neural representation of $U2$ attends to H_{RNQ} and $U10$ attends to H_{RQ} simultaneously. In other words, Q and NQ "mutually" attend to each other's history. Mathematically, as for Q and H_{RNQ}: the Q containing historical information can be obtained via soft alignment of words in the question $Q = [h_1^q, h_2^q, ..., h_N^q]$ and history turns $H_{RNQ} = \{d_1^p, d_2^p, ..., d_B^p\}$ as follows (see Part I in Figure 2):

$$u_i^q = [h_i^q, c_i^q] \tag{3}$$

where $c_i^q = att(h_i^q, H_{RNQ})$ is an attention-pooling vector of the whole history(H_{RNQ}), , and v and W are the weights:

$$\begin{aligned} s_j^i &= v^T tanh(W_Q h_i^q + W_H d_j^p) \\ a_k^i &= exp(s_k^i) / \sum_{j=1}^{B} exp(s_j^i) \\ c_i^q &= \sum_{k=1}^{B} a_k^i d_k^p \end{aligned} \tag{4}$$

Each word representation in Q dynamically incorporates aggregated matching information from the history H_{RNQ}.

The final representations of the question and the non-question after mutual attention are $Q' = [u_1^q, u_2^q, ..., u_N^q]$ and $NQ' = [u_1^p, u_2^p, ..., u_M^p]$. Each word vector in Q' and NQ' not only represents the original sentence meaning but also contains dialogue context. The effects of different choices of turns in history and the ways of aggregating the history will be discussed later.

3.3 Match-LSTM

We follow the work of Wang and Jiang [34] and adopt match-LSTM to capture the features between the processed Q' and NQ' word by word.

As Part II of Figure 2 shows, a one-layer LSTM is used to encode the representations with mutual attention for questions and non-questions. We thus obtain $Q'' = \{f_i^q\}_{i=1}^N$ and $NQ'' = \{f_i^p\}_{i=1}^M$. When encoding the non-question, we introduce a series of attention-weighted combinations of the hidden states of the question, where each combination is for a particular word in the non-question. The sentence-pair representation $P = \{p_i\}_{i=1}^M$ is calculated with attention mechanism as follows.

$$p_i = LSTM(p_{i-1}, [f_i^p, c_i]) \tag{5}$$

where $c_i = att(Q, f_i^p, p_{i-1})$ is an attention-pooling vector of the whole question(Q):

$$\begin{aligned} s_j^i &= v^T tanh(W_{NQ} f_i^p + W_Q f_j^q + W_p p_{i-1}) \\ a_k^i &= exp(s_k^i) / \sum_{j=1}^{N} exp(s_j^i) \\ c_i &= \sum_{k=1}^{N} a_k^i f_k^q \end{aligned} \tag{6}$$

Finally, we use p_M to represent the whole Q-NQ pair which is used for predicting the final result.

3.4 Prediction Layer

At the last step, we use a fully-connected (FC) layer with softmax to do binary classification, which indicates whether this pair of utterance has QA relation.

Driven by the intuition that distance is a critical feature for matching QA pairs, we explicitly add the distance at the end of our model to preserve its effectiveness. The distance d is defined as a 10-dimensional one-hot vector, where the position of 1 is equal to the distance. For example, if the distance is 4, then the 4th dimension of the vector is set to 1. If $d \geq 10$, then the 10th dimension is set to 1.

Finally, the probability is calculated as follows:

$$\begin{aligned} FC &= W[p_M, d] + b \\ P(Q, NQ) &= Softmax(FC) \end{aligned} \tag{7}$$

In sum, the input of our model is a Q and NQ with associated information (history and distance), and the output of model is the probability of being *True* or *False* which indicates if the Q is the matched question of the NQ. Hence, the loss function is the cross entropy between the predicted probability $P(Q, NQ)$ and the ground truth. The model can be seen as a binary classification model and all the parameters are trained altogether except the pretrained word embeddings.

3.5 Greedy QA Matching

Based on the trained model, every NQ now has a matching probability with every Q before it in the dialogue sequence. The greedy algorithm matches the NQ with the Q with maximum probability if that probability exceeds 0.5. The threshold is set to 0.5 because our model is actually a two-class classifier.

4 Dataset Construction

Previous QA datasets are in the form of independent QA pairs [38] and do not provide surrounding dialogue context. Although He et al.[18] solved a similar problem, their customer service dataset is not open to public due to privacy concerns. Wei et al. [36] published their dialogue dataset collected from an online forum. However, their work focuses on the dialogue policy learning and the data doesn't preserve the original utterances.

There is also no published qualified dialogue dataset for the QA matching task. The IRC dataset [13] and Reddit dataset [20] are both multi-party dialogues instead of two-party dialogues. Twitter Triple Corpus [29] and Sina Weibo [27] are not multi-turn dialogues with only two or three turns. MultiWOZ 2.0 [4], CamRest676 [26], and Stanford Dialog Dataset [14] are multi-turn dialogues with dialogue act annotations but these QA pairs appear next to each other. If we shuffle the well-ordered dataset randomly or by some rules, it's unnatural and incorrect because the original dialogue context is destroyed as shown in Table 1. The two-party Ubuntu dataset [23] meets these requirements and we annotated 1000 dialogue sessions. However, the statistics and experiments show that the Ubuntu dataset is not a good QA matching dataset especially for long distance QAs, given the limited manpower for annotation. More details will be shown in Section 6.

Turn	QA	Original	Shuffled
r1:Where can I enjoy my holiday? I want to go somewhere near the sea and warm.	Q1	1	1
r2:Maybe *Xiamen* is a good choice.	A1	2	4
r1:Is *there* anything delicious?	Q2	3	2
r2:Wow, that's quite a lot. There are......	A2	4	3

Table 1. An example of shuffling well-ordered dialogues. Unreasonable co-reference appears after shuffling.

Hence, we create a new dataset suitable for this task. Nearly 160,000 distinct dialogues are collected from an online health forum[8].All the personal information was removed in advance by the website. After some basic data cleaning methods such as deleting the irrelavant sentences like "Please pay ** coins to continue consultation", we labeled 1000 randomly selected two-party multi-turn dialogues with Q (question), A (answer) and O (others) labels. A small amount of turns (0.24%) are considered by the annotators to be both a question and an answer, and these are treated as questions uniformly. The Fleiss Kappa between three annotators was 0.75, indicating substantial agreement.

On average, each dialogue has 19.68 doctor turns and 17.32 patient turns. Most turns are made up of a sentence fragment, so the number of words for each turn is on average less than 10 words [9]. 21.9% of the questions have no answers, 22.7% of the questions have more than one answer and the remaining questions have the only answer. For questions that do have answers, each of them is matched to 1.41 answers on average.

The annotated dialogues are split into training/development/test sets by 7:1:2. The distribution of the QA pairs by distance is shown in Table 2.

Distance Dataset	1	2	3	4	≥ 5
Training	3439	2068	1029	450	554
Development	454	331	167	76	99
Test	947	592	274	136	168

Table 2. The distribution of QA pairs by Q-A distances.

We reconstructed the labeled dialogues into Q-NQ pairs with distance, history and binary golden label used for our models. A NQ from a party is paired with every earlier Q from the other party. If the pair is a QA pair, it is labeled as True(T). Otherwise, it is labeled as False(F). The distribution of positive and negative data in three datasets is shown in Table 3.

Dataset Label	Training	Development	Test
True	7540	1226	2116
False	80631	14889	23893

Table 3. The distribution of positive and negative Q-NQ pairs on three datasets.

5 Experiment Setup

In this section, we first list the baselines and the ablations of our full model. Next, we define the evaluation metrics. Finally, we show the details of hyperparameters in our model.

[8] An dialogue example online: `https://www.120ask.com/shilu/0cjdemaflhj6oyw9.html`

[9] There are on average 9.80 questions with 8.89 words, 10.78 answers with 6.62 words, 16.41 casual chit chats with 6.99 words in each session according to annotated dialogues.

5.1 Baselines and Our Method

5.1.1 Baselines

We mainly have the following four types of baselines.

- **Greedy strategy (GD)** A simple baseline *Greedy* is that, when a question is posed by RQ, we can directly match the following several NQs said by $\overline{RQ}$ as the answers. It stops when meeting another Q or a turn said by RQ. There are a few variants in this methods. GD1 selects only one satisfying answer, and GDN selects multiple satisfying answers. The methods with *Jump* (**J**) skip the non-question sentence uttered by RQ when matching the NQs.
- **Distance** A simple model takes a 10-dimensional one-hot distance vector as the input of a fully-connected layer and outputs the score for each Q-NQ pair.
- **Word-by-word match LSTM (mLSTM)** This model is proposed by Wang et al. [34], used for natural language inference. It performs word-by-word matching based on an attention mechanism, with the aim of predicting the relation between two sentences.
- **Recurrent Pointer Network (RPN)** The model proposed by He et al. [18] is the previous state-of-the-art method for a similar task, and is implemented with some modifications to fit our task. We use two parallel RPN to distinguish questions from two parties. Comparing the classification loss and regression loss proposed in this paper, we choose the one that performs best on our test set.

5.1.2 Our Models

By disabling some of its components, our model comes in three main variants:

- **Distance Model (DIS)** removes the mutual attention. It directly puts Q and NQ with the distance into Part II of the full model in Figure 2.
- **History Model (HTY)** disables the distance information at the prediction layer.
- **History-Distance Model (HDM)** is the full model we have explained in Section 3.

5.2 Evaluation Metrics

Once we have identified all of the QA pairs, we count the true positive, false positive and false negative for each question. To measure the quality of the matched QA pairs, micro-averaging precision (**P**), recall (**R**) and F1-score (**F1**) are calculated, with all questions in test dataset treated equally. Another metric, accuracy (**Acc**), is used when evaluating QA pairs matched with a specific distance. Accuracy is calculated by the percentage of ground truth QA pairs that are predicated positive by the models.

5.3 Implementation Details

We use Jieba[10] to do Chinese word segmentation and pre-train the 100 dimentional word embeddings with Skip-gram model [24] on all of the dialogues (including the unlabeled ones) we have collected online. For our proposed models, we use LSTM with hidden state size of 128 and 256 for Part I and Part II of the model respectively. We adopt Adam optimizer with 0.001 learning rate and 0.3 dropout. Learning rate decay is 0.95 and the training process terminates if the loss stops reducing for 3 epochs. All experimental results are averaged over three runs.

6 Results and Analysis

In this section, we show the end-to-end results and ablation tests for the specific architectural decisions.

6.1 Overall Performance

The main results of different models are shown in Table 4. The last row lists the human performance, regarded as the upper bound of this task. Note that human performance is not perfect due to inherent ambiguities in the dialogues.

Models	P	R	F1
GD1	69.84	44.73	54.53
GDN	70.03	69.11	69.57
GD1+J	70.38	50.40	58.74
GDN+J	51.47	82.90	63.51
mLSTM	58.17	4.20	7.84
Distance	71.57	69.34	70.44
RPN	72.40	68.63	70.46
DIS	78.46⋆	70.34	74.70⋆
HTY	75.40⋆	76.42⋆	75.90⋆
HDM	76.44⋆	78.44⋆	**77.43**⋆
Human	85.11	84.21	84.66

Table 4. The end-to-end performance of all methods on test dataset. Scores marked with ⋆ are statistically significantly better than the RPN with $p < 0.01$.

The results of the rule-based methods are not bad, which indicates that questions are actually followed by their answers in many cases. The GDN increases the F1-score to 69.57% compared with Greedy-1 because it can solve the case of simple fragmented answers. For GDN+J, the recall is the best among all the methods while accuracy and F1-score suffer. The reason is that GDN tends to match NQ with Q as much as possible, so many chit chats will be regarded as answers, which reduces the precision.

Model mLSTM underperforms because it is difficult to solve the QA matching problem with only two short texts without history. The word distribution between the questions and answers are quite different and maybe unrelated without background knowledge. Distance achieves good scores which shows that the distance is very important factor when identifying QA relations in dialogues. People tend to answer a question the moment they see it except in the case of incremental QAs. RPN obtains competitive results. It mainly benefits from taking the dialogue session as a whole which contains all the information in a session.

Our proposed models achieve the best results compared with above models. The HDM improves the F1-score to 77.43%, significantly better than RPN by t-test with $p < 0.01$. Although the recall of HDM is not better than GDN+J and the precision is lower than DIS, the overall quantity and quality of QA pairs identified are the best, shown by the highest F1-score. In addition, the comparable results achieved by HTY demonstrate that QA matching not only depends on the distance but also relies on the history information. This shows that HDM model successfully combines both the distance and history information.

6.2 Variable Distance Matching

According to the results above, we can find that the model Distance, RPN, DIS, HTY and HDM are competitive. Thus, we further analyze the accuracy of these five models on variable distances.

[10] https://github.com/fxsjy/jieba

Models	1	2	3	4	≥ 5
Distance	**100.0**	88.01	0.0	0.0	0.0
RPN	89.37	69.37	50.12	36.96	13.10
DIS	96.23	**89.13**⋆	17.03	2.45	0.0
HTY	94.37	78.89⋆	57.42	38.48	**28.17**⋆
HDM	95.99	83.16⋆	**59.37**⋆	**40.44**	24.80⋆

Table 5. The Acc (%) of matched QA pairs on variable distances.

Table 5 shows that Distance and DIS work well on SQAs but deteriorate rapidly as the distance between QAs grows, indicating that relying solely on the distance information is insufficient. On the other hand, using dialog context or history, the matching accuracy of RPN and HTY is generally lower on SQAs but higher on LQA. Our full model (HDM) actually is a good trade-off by incorporating both distance and history information. This also accords with the decision process made by human annotators.

6.3 Ablation Tests

We justify the design of our model in the following two aspects.

6.3.1 Different definitions of history

To show the effectiveness of using the turns between Q and NQ as history, we devise the following variants on the HDM model for comparison:

- **Q-history Model (QH)** has the same structure of HDM where the history is all the turns before Q.
- **A-history Model (AH)** has the same structure of HDM where the history is all the turns before NQ.

The main results with different choices of history are shown in Table 6. Our final model (HDM) outperforms QH and AH, indicating that the turns between Q and NQ are significant when figuring out the relation of Q-NQ pair. The turns before Q is actually not that important for matching Q and NQ. Although there is an overlap between the history we defined and the turns before NQ, the turns before Q brings more noises than benefits for the end-to-end performance. This suggests that our definition of history as the turns between Q and NQ is reasonable and effective.

Models	F1	Acc@3	Acc@4	Acc@≥ 5
QH	74.56	21.53	21.53	0.79
AH	73.84	15.81	10.78	4.76
HDM	**77.43**	**59.37**	**40.44**	**24.80**

Table 6. The matching results on different choices of history.

6.3.2 Different ways of attending to the history

To evaluate the effectiveness of mutual attention for aggregating the history, we devised variants of the HDM model for comparison:

- **Non-mutual Model (NM)** has the same structure of HDM where Q attends to H_{RQ} and NQ attends to H_{RNQ}.
- **Identical history model (ID)** has the same structure of HDM where Q and NQ attends to the same history $H_{RQ} \bigcup H_{RNQ}$.

The main result in Table 7 reveals that our choice of separating the history by role label and mutually attending to each other does work. The full model (HDM) consistently outperforms both NM and ID.

For the ablation test on different ways of attending to the history, we conclude that NM is better than ID. It's due to better understanding on individual speakers which can help the understanding of the dialogue, similar to the idea in the AAAI2019 paper "A Deep Sequential model for Discourse Parsing on Multi-party Dialogues". Our work targets the QA relations in dialogues which are more related to the interactions between speakers. As a result, our full model is better than NM.

The difference between our full model HDM and ID is that in ID both Q and NQ attend to all turns in the history regardless of who uttered those turns, whereas HDM employs a mutual attention mechanism that distinguishes turns by their speakers. Specifically, the Q only attends to those turns uttered by the NQ speaker, while the NQ attends to those turns by the Q speaker. This resembles to some extent the firm attention in Amplayo's paper[1]. This will help the model to focus on the interactions between speakers.

We also conduct a Z-test on the results to show that the improvements on the LQAs are statistically significant even with a small sample size.

Models	F1	Acc@3	Acc@4	Acc@≥ 5
NM	75.81	57.18	37.50	20.04
ID	75.46	54.50	28.43	11.70
HDM	**77.43**	**59.37**	**40.44**	**24.80**

Table 7. The results on different ways of aggregating history.

6.3.3 Example Outputs

To provide a better understanding of the behavior of our models, we include an example output in Table 8. It contains both LQAs and SQAs. In this case both HYD and HDM predicts QA relations better than the baseline RPN. As for SQAs, all of the models perform well. However, the DIS model is obviously not capable of matching LQA pairs. It indicates that the distance information sometimes hurts the performance of HDM on matching LQA pairs.

6.4 Results on the Ubuntu Dialogue Corpus

The Ubuntu dataset [23] is a large unannotated dialogue corpus designed for next utterance classification and dialogue generation. It is extracted from multi-party dialogues, while conversation disentanglement is still an unsolved problem. Therefore, many dialogues in this dataset don't make much sense because the two parties are not actually talking to each other (they were talking to a third party whose utterances were removed). It maybe a good dialogue resource with 930,000 dialogues, but it is impossible to label such large amount of dialogues to filter out all the irrelevant turns.

We randomly sampled 1000 dialogues each with more than 10 turns. Then we annotated the extracted corpus with (Q, A, O) labels as mentioned above. If the current dialogue contains no Q-NQ pairs, it is replaced by a new dialogue with similar number of turns. Finally, we collected 1000 annotated dialogues. The QAs in Ubuntu

Ground Truth	RPN	DIS	HTY	HDM	Role	Utterances
Q1					P	Boy, 4 months. He tried a little yolk yesterday and shat that night but haven't shat until today what's wrong????
O	O	O	O	O	D	Hello
O	O	O	O	O	P	Hello
Q2					D	Is he four months old
A2	A2	A2	A2	A2	P	Yes
A1	A1	O	A1	A1	D	Eat too early
A1	O	O	A1	A1	D	Not advise
A1	O	O	A1	A1	D	Difficult for digestion

Table 8. A correct case of predictions and human annotations in our dataset.

corpus mainly focus on step-by-step operations and the dialogues are lack of long distance QAs. The results are showed in Table 9.

Models	P	R	F1
GD1	89.33	58.96	71.03
GDN	84.62	78.86	81.64
GD1+J	83.62	65.02	73.16
GDN+J	57.05	89.49	69.68
mLSTM	64.88	0.37	0.73
Distance	84.97	78.99	81.87
RPN	71.33	61.80	66.23
DIS	85.36	66.21	74.57
HTY	85.25	77.53	81.20
HDM	85.88	77.53	81.48
Human	90.28	84.42	87.25

Table 9. The end-to-end performance of all methods on the Ubuntu test dataset.

The results of Distance and GDN achieves the top-2 highest F1 score, which is consistency with the fact that 41.95% of questions have no answers and 79.55% of the QA pairs are consecutive in the dialogue. However, it still shows the effectiveness of our full model HDM, which achieves the competitive results to the best results.

7 Related Work

Detection of QA pairs from online discussions has been widely researched these years. Shrestha and Mckeown [28] learned rules using Ripper for detecting QA pairs in email conversations. Ding et al. [10], Kim et al. [21] and Catherine et al. [6] applied the supervised learning method including conditional random field and support vector machine. Cong et al. [8] proposed an unsupervised method combining graph knowledge to solve the task. Catherine et al. [5] proposed semi-supervised approaches which require little training data. He et al. [18] used the pointer network to find QA pairs in Chinese customer service. However, the tasks mentioned above are all different from ours. We identify QA pairs from two-party dialogues on online discussion forum, and focus especially on long-distance QA pairs. Besides, our dialogue is constrained between two roles who can both utter questions and answers.

There exists several methods in other tasks which can be adapted to our QA matching problem. Feature-based method is popular for solving many NLP problems. In the work of Ding et al. [10], Wang et al. [33] and Du et al. [12], they examined lexical and semantic features in two sentences for QA matching. However, the features such as common question words and roles have already been explicitly annotated in our data. Besides, other features such as special word occurrence or time stamp are unavailable here. According to the data, we considered the distance as the most important feature and implemented this feature-based method as one baseline. Recent researches using deep neural networks have increased a lot. He and Lin [17] and Liu et al. [22] used the sentence pair interaction approach which takes word alignment and interactions between the sentence pair into account. Attention mechanism was also added for performance improvement [25, 34, 7]. We also use word alignment and interactions to calculate the QA similarity. Specially, we adopt attention mechanism to solve the LQA cases.

There are other kinds of alignment problems such as temporal sequences alignment. Video-text alignment is one of the temporal assignment or sequence alignment problems. Previous work automatically provides a time (frame) stamp for every sentence to align the two modalities such as [3] and [11]. Bojanowski et al. [3] extended prior work by including the alignment of actions with verbs and aligned text with complex videos. Dynamic time warping (DTW) is anothor algorithm for measuring similarity between two temporal sequences. It's also widely used in video-text alignment task [11], speech recognition task [31].

8 Conclusion

In this paper, we focus on identifying QA pairs in two-party multi-turn online dialogues based on turns with Q or NQ labels. Our proposed models achieve the best accuracy overall, and perform particularly well on LQAs. We also discuss the model decisions of using two attention mechanisms in an mutual way and the definition of history. Our future work will focus on more discourse relations in dialogues. Utilizing out-of-domain knowledge is another research direction for utterances matching task, especially for health-related dialogues.

REFERENCES

[1] Reinald Kim Amplayo, Seonjae Lim, and Seung-won Hwang, 'Entity commonsense representation for neural abstractive summarization', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 697–707, (2018).

[2] Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos, 'Discourse structure and dialogue acts in multiparty dialogue: The stac corpus', (2016).

[3] Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid, 'Weakly-supervised alignment of video with text', in *Proceedings of the IEEE international conference on computer vision*, pp. 4462–4470, (2015).

[4] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic, 'Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, (2018).

[5] Rose Catherine, Rashmi Gangadharaiah, Karthik Visweswariah, and Dinesh Raghu, 'Semi-supervised answer extraction from discussion forums', in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1–9, (2013).

[6] Rose Catherine, Amit Singh, Rashmi Gangadharaiah, Dinesh Raghu, and Karthik Visweswariah, 'Does similarity matter? the case of answer extraction from technical discussion forums', in *Proceedings of COLING 2012: Posters*, pp. 175–184, (2012).

[7] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen, 'Enhanced lstm for natural language inference', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1657–1668, (2017).

[8] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun, 'Finding question-answer pairs from online forums', in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 467–474. ACM, (2008).

[9] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou, 'Superagent: A customer service chatbot for e-commerce websites', in *Proceedings of ACL 2017, System Demonstrations*, pp. 97–102, (2017).

[10] Shilin Ding, Gao Cong, Chin-Yew Lin, and Xiaoyan Zhu, 'Using conditional random fields to extract contexts and answers of questions from online forums', in *Proceedings of ACL-08: HLT*, pp. 710–718, (2008).

[11] Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross, 'A neural multi-sequence alignment technique (neumatch)', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8749–8758, (2018).

[12] Wenchao Du, Pascal Poupart, and Wei Xu, 'Discovering conversational dependencies between messages in dialogs', in *Thirty-First AAAI Conference on Artificial Intelligence*, (2017).

[13] Micha Elsner and Eugene Charniak, 'You talking to me? a corpus and algorithm for conversation disentanglement', in *Proceedings of ACL-08: HLT*, pp. 834–842, (2008).

[14] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning, 'Key-value retrieval networks for task-oriented dialogue', in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 37–49, (2017).

[15] Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters, 'Modelling and detecting decisions in multi-party dialogue', in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 156–163. Association for Computational Linguistics, (2008).

[16] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins, 'Learning to forget: Continual prediction with lstm', (1999).

[17] Hua He and Jimmy Lin, 'Pairwise word interaction modeling with deep neural networks for semantic similarity measurement', in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 937–948, (2016).

[18] Shizhu He, Kang Liu, and Weiting An, 'Learning to align question and answer utterances in customer service conversation with recurrent pointer networks', (2019).

[19] Zongcheng Ji, Zhengdong Lu, and Hang Li, 'An information retrieval approach to short text conversation', *arXiv preprint arXiv:1408.6988*, (2014).

[20] Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang, 'Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1812–1822, (2018).

[21] Su Nam Kim, Li Wang, and Timothy Baldwin, 'Tagging and linking web forum posts', in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 192–202. Association for Computational Linguistics, (2010).

[22] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, 'Modelling interaction of sentence pair with coupled-lstms', *arXiv preprint arXiv:1605.05573*, (2016).

[23] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau, 'The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems', in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294, (2015).

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *Workshop Track Proceedings of the 1st International Conference on Learning Representations*, (2013).

[25] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom, 'Reasoning about entailment with neural attention', *arXiv preprint arXiv:1509.06664*, (2015).

[26] Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke, 'A network-based end-to-end trainable task-oriented dialogue system', in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain*, pp. 438–449, (2017).

[27] Lifeng Shang, Zhengdong Lu, and Hang Li, 'Neural responding machine for short-text conversation', in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1577–1586, (2015).

[28] Lokesh Shrestha and Kathleen McKeown, 'Detection of question-answer pairs in email conversations', in *Proceedings of the 20th international conference on Computational Linguistics*, p. 889. Association for Computational Linguistics, (2004).

[29] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan, 'A neural network approach to context-sensitive generation of conversational responses', in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 196–205, (2015).

[30] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer, 'Dialogue act modeling for automatic tagging and recognition of conversational speech', *Computational linguistics*, **26**(3), 339–373, (2000).

[31] Taras K Vintsyuk, 'Speech discrimination by dynamic programming', *Cybernetics and Systems Analysis*, **4**(1), 52–57, (1968).

[32] Oriol Vinyals and Quoc Le, 'A neural conversational model', *arXiv preprint arXiv:1506.05869*, (2015).

[33] Baoxun Wang, Xiaolong Wang, Chengjie Sun, Bingquan Liu, and Lin Sun, 'Modeling semantic relevance for question-answer pairs in web social communities', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1230–1238. Association for Computational Linguistics, (2010).

[34] Shuohang Wang and Jing Jiang, 'Learning natural language inference with lstm', in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1442–1451, (2016).

[35] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou, 'Gated self-matching networks for reading comprehension and question answering', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 189–198, (2017).

[36] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai, 'Task-oriented dialogue system for automatic diagnosis', *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 201–207, (2018).

[37] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li, 'Building task-oriented dialogue systems for online shopping', in *Thirty-First AAAI Conference on Artificial Intelligence*, (2017).

[38] Yi Yang, Wen-tau Yih, and Christopher Meek, 'Wikiqa: A challenge dataset for open-domain question answering', in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, (2015).