What did you refer to? Evaluating Co-references in Dialogue

Weinan Zhang[♯], Yue Zhang[♭], Hanlin Tang[♢], Zhengyu Zhao[♯], Caihai Zhu[♯], Ting Liu[♯]
[♯]Harbin Institute of Technology, Harbin, China
[♭]Westlake University, Hangzhou, China
[♢]School of Computer Science and Technology, Beijing, China
{wnzhang, zyzhao, chzhu, tliu}@ir.hit.edu.cn
zhangyue@westlake.edu.cn
hltang@bit.edu.cn

Abstract

Existing neural end-to-end dialogue models have limitations on exactly interpreting the linguistic structures, such as ellipsis, anaphor and co-reference, etc., in dialogue history context. Therefore, it is hard to determine whether the dialogue models truly understand a dialogue or not, only depending on the coherence evaluation of their generated responses. To address these issues, in this paper, we proposed to directly measure the capability of dialogue models on understanding the entity-oriented structures via question answering and construct a new benchmark dataset, DEQA, including large-scale English and Chinese human-human dialogues. Experiments carried on representative dialogue models show that these models all face challenges on the proposed dialogue understanding task. The DEQA dataset will release for research use.

1 Introduction

Driven by the growth of interest in social chatbot, online customer service and virtual mobile assistant, social dialogue systems have received increasing research attention (Cui et al., 2017; Zhou et al., 2018; Hancock et al., 2019). The current dominant method has been sequence-to-sequence models, trained over large dialogue data end-to-end. Such models use neural network architectures such as Transformer (Vaswani et al., 2017) to encode a user utterance and a dialogue history before generating a system utterance (Adiwardana et al., 2020; Roller et al., 2020; Bao et al., 2020). A major advantage is the use of standard and general model architecture, which facilitates end-to-end training process over large scale dialogue text (Shang et al., 2015; Zhang et al., 2018, 2019, 2020).

Dialogue		
U_1:	Well, you know how important a clean house is to your grandma.	
U_2:	Yes, I hear about it every time she comes here.	
	What do you hear about?	Q_1
	A clean house.	A_1
U_1:	She was the head janitor at St. Mary's Hospital for thirty years, after all.	
U_2:	I think she misses that job and wants to take it out on us.	
U_1:	You know, maybe she is just a neat freak.	
	Who is just a neat freak?	Q_2
	Grandma.	A_2
U_2:	I think she just likes to make us miserable.	
U_1:	You could be right.	

(a)

Co-reference Chain (OntoNotes style)	
Chain 1 (IDENT)	Chain 3 (IDENT)
1.6-8 a clean house	1.12-12 grandma
2.5-5 it	2.8-8 she
Chain 2 (IDENT)	3.1-1 she
3.4-9 head janitor at St. Mary's Hospital	4.3-3 she
	5.4-4 she
4.5-6 that job	6.3-3 she

(b)

Table 1: (a) Sample of English dialogue in the proposed dataset. U_1 and U_2 are two interlocutors in the dialogue. Q_i and A_i (i=1,2) are clarification requests and the corresponding answers. (b) Co-reference chain annotation in OntoNotes 5.0 style.[1]

Despite showing effectiveness in empirical evaluation, existing work has a few important limitations. First, it is difficult to visualize or interpret the representation of dialogue state from a dense neural network encoder. In particular, there is not explicit representation of entities, semantic relations or discourse structures. Second, the performance of a dialogue system is evaluated directly by the quality of the generated responses. However, relatively little work has been done on evaluating how a system response is determined, which can be important because a proper response can be generated by simply relying on superficial and spurious patterns in dialogues, and we want to find out the cause of

[1]IDENT denotes the entities in a co-reference chain are identical. "1.6-8" indicates that "a clean house" is from the 6th to 8th tokens in the 1st utterance.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5075–5084

problematic responses for identifying model limitations. ; To address such limitations, it can be useful to directly measure the quality of *dialogue understanding* by asking a dialogue model to identify important structures in dialogue histories. In this paper, we focus on entity level understanding, evaluating references to entities in a dialogue history context. Such references can include explicit anaphora and implicit mentions by using zero pronouns. Take Table 1 (a) as an example, where the dialogue history consists of 7 utterances and the second utterance contains a pronoun "*it*". At this point, we can measure system understanding of the dialogue state by checking whether the system can resolve the anaphora concerning "*a clean house*".

Our goal is to provide a large-scale benchmark and to evaluate the performance of social chatbot systems on dialogue understanding concerning entities. One way to define the task is to cast it as a co-reference resolution problem (Yin et al., 2017; Kong et al., 2019; Quan et al., 2019), where a benchmark can be constructed by manually labeling co-reference information on a dialogue dataset, as shown in Table 1 (b). However, such a benchmark does not fully meet our goal because a separate model is necessary for achieving co-reference resolution, and it may be challenging to seamlessly integrate such a co-reference module into a dialogue model being tested.

We take a different method, checking dialogue understanding of dialogue systems by inserting clarification requests (Schlangen, 2004; Stoyanchev and Johnston, 2015) into dialogues, and evaluating the response of dialogue systems on such requests. One example is shown in Table 1 (a), where we break a dialogue in the middle, adding clarification requests. For example, for the question "*Who is just a neat freak?*", the correct system response should be "*Grandma*", which reflects that the model has correct understanding of the dialogue context.

The advantage is three fold. First, this method allows the evaluation of a dialogue system without using an external probe task, by directly evaluating system generated responses. This makes our benchmark directly useful for evaluating arbitrary social dialogue models. In contrast to open-ended responses in chit-chats, responses for the proposed clarification requests are factual thus facilitating automatic evaluation. Second, it allows easier crowdsourcing for dataset construction as compared with co-reference resolution, which requires strict training of manual labelers for understanding linguistic concepts. It is thus useful for acquiring large-scale datasets. Such observation is consistent with recent work on other NLP tasks (FitzGerald et al., 2018; Roit et al., 2020). Third, this method allows easy extension to dialogue understanding beyond the entity reference level, such as event co-references, semantic relations and discourse level understanding. No new labeling standards are necessary for adding a new task.

According to the above observations, we create a large scale benchmark, open domain Dialogue Entity via Question Answering (DEQA), which consists of one English dataset and one Chinese dataset, of 8,415 and 6,203 dialogues, respectively. Each dialogue contains one or more questions similar to the one in Table 1. We choose to evaluate representative multi-turn neural dialogue systems, including models using Transformer (Vaswani et al., 2017) and DialoGPT (Zhang et al., 2020). Results show that the prevalent models of multi-turn dialogue generation face challenges in the co-reference questions. We will release the dataset at Github [2] for research use.

2 Dataset

We present the task (Section 2.1), the linguistic structures to evaluate (Section 2.2), the dataset construction (Section 2.3), the dataset characteristics (Section 2.4) and the evaluation metrics (Section 2.5) below.

2.1 Task Definition

Given a multi-turn dialogue, the task is to answer questions concerning one or more turns of the dialogue history. In particular, the model needs to answer questions about the anaphor and ellipsis phenomenons that appear in the context. It is worth noting that most of the answers can be extracted from the given context, but some answers may not explicitly appear in the context. These questions are called *summary questions*. The dialogue model should also have the capability on answering these summary questions.

We have already seen one example of English dialogue in Table 1. Table 2 shows a sample Chinese dialogue in the annotated dataset and its English translation. For the second utterance "我也想吃。。。(I also want to eat...)", the corresponding question is "你也想吃什么？(What do

[2] https://github.com/adamszhu/DEQA

	Chinese Dialogue	English Translation
U_1:	我想吃炸鸡。。。	I want to eat fried chicken...
U_2:	我也想吃。。。	I also want to eat... (literal translation)
U_1:	昨天买的炸鸡被我家猫吃了。	The fried chicken I brought yesterday was eaten by my cat.
U_2:	哈哈，是时候教育它了	Haha, it's time to teach it a lesson.
U_1:	不舍得啊	Unwilling to do that.
U_2:	我来替你啊	I can do it for you.

Table 2: Sample of Chinese dialogue in the annotated dataset.

you want to eat too?)" This question refers to the first utterance and the phrase "炸鸡(fried chicken)" should be extracted as the answer of the question. For the fourth utterance "哈哈，是时候教育它了。(Haha, it's time to teach it a lesson.)", there is a pronoun "它(it)" which should be resolved. A question "教育谁？（Teach whom a lesson?）" which refers to the fourth utterance is then raised. According to the third utterance, the answer of the question is "你的猫(your cat)". Note that in the proposed task, some answers should be summarized from the whole dialogue rather than only one utterance.

2.2 Linguistic Structures

In our dataset, a question is raised towards one ellipsis or anaphor phenomenon in a given dialogue. The role of a raised question can be seen as a "label" of a pronoun or zero pronoun. Correspondingly, the answer to the question is thus the antecedent of the pronoun or zero pronoun. Ellipsis, anaphor and co-reference are used frequently in natural language expression, especially in human-human dialogues (Quan et al., 2019). The examples in dialogues include:

- Ellipsis
  1) Zero anaphora (Noun phrase ellipsis)
  U_1: "你喜欢邦乔维的音乐吗？" ("*Do you like music of Bon Jovi?*")
  U_2: "是的(*Yes*)，我(*I*)喜欢(*like*)。"

The noun phrase "邦乔维的音乐" ("*music of Bon Jovi*") is omitted in the second utterance.

2) Verbal phrase ellipsis
U_1: "*I like the V6 engine of Audi S4.*"
U_2: "*So do I.*"

Here, "*do*" is a trigger word which indicates the ellipsis of verbal phrase "*like the V6 engine of Audi S4*".

- Anaphor
  1) Personal pronoun
  U_1: "*Do you know Kelly Clarkson?*"
  U_2: "*Yes, she is my idol.*"

"*she*" is a personal pronoun which refers to "*Kelly Clarkson*".

2) Demonstrative pronoun
U_1: "*Have you ever made some family albums?*"
U_2: "*Yes, these are my treasures.*"

Here, "*these*" refers to "*family albums*".

- co-reference
  U_1: "*There is a concert of Taylor Swift next month.*"
  U_2: "*Let us sing with Swifty together!*"

In this case, "*Swifty*" and "*Taylor Swift*" are co-reference.

2.3 Data Annotation

The English dialogue data are sourced from the DailyDialogue dataset (Li et al., 2017). The Chinese dialogue data are collected by ourself from Douban[3], a Chinese online forum. We randomly sample a subset of dialogues from the above two dataset respectively, and then annotate these dialogues in question answering.

For annotation, the first step is to identify ellipsis, anaphor and co-reference phenomenons of utterances in dialogue data. For each utterance, annotators determine whether the meaning of the utterance is complete when ignoring the dialogue context. If the meaning of an utterance is determined as incomplete, we can further identify ellipsis. Both zero anaphora and verbal phrase ellipsis are determined, and the anaphor can include both the personal and demonstrative pronouns. However, utterances such as "是(Yes)", "不是(No)", "好的(OK)" are not identified as ellipsis.

The second step is to raise questions to the identified zero anaphora, personal pronouns, demonstrative pronouns and corefered entities. We require that the questions are not simply obtained

[3] https://www.douban.com/

	English	Chinese
Total # of dialogues	8,415	6,203
Total # of questions	11,904	10,387
Training set	6,603	4,962
Dev set	832	621
Test set	830	620

Table 3: Statistics of the annotated dialogue dataset. # denotes "number".

	English		Chinese	
	#	%	#	%
Entity	4,765	40.03	5,810	55.94
Phrase	2,298	19.30	2,203	21.21
Clause	87	0.73	755	7.26
Fragment	4,754	39.94	1,619	15.59

Table 4: Statistics of the answer types in the annotated dataset. # denotes "number" and % denotes percentage.

	English		Chinese	
	#	%	#	%
Seq	11,055	92.87	9,831	94.65
Skip	145	1.22	206	1.98
Cross	146	1.23	195	1.88
Summary	552	4.68	147	1.49

Table 5: Statistics of the status of an answer that appears in a dialogue. # denotes "number" and % denotes percentage.

by adding interrogative words in the original utterances. The third step is to give answers of each question, which may be an entity, a phrase, a chunk, a clause or a fragment of an utterance. Table 3 presents the statistics of the annotated dialogue dataset. In the end, we have 11,904 questions being labeled in 8,415 English dialogues, and 10,387 in 6,203 Chinese dialogues.

2.4 Characteristics

The characteristics of the annotated dialogue dataset include:

1) The dialogues are real human-human conversations;

2) Each dialogue is annotated with one or more questions and each question is related to at least an utterance of a dialogue;

3) The answer of a question may appear in one or more utterances of the dialogue. It means that an answer may be a composition of fragments that are from different utterances rather than a continuous span of an utterance. We analyze the types of answers in the annotated dataset below.

First, Table 4 shows the number and percentage of different types of answers.

We can see that most of the answers are entity, phrase and fragment. The proportion of the clause type in the Chinese dialogue data is about ten times that in the English dialogue data. The proportion of the fragment type in English is larger than that in Chinese.

Second, we give the statistics of the status of an answer that appears in a dialogue. Here, in Table 5, "Seq" and "Skip"denote the tokens of an answer are sequential and skipping in an utterance, respectively. "Cross" indicates that the tokens of an answer are from different utterances. "Summary" means that the tokens of an answer written by the annotator is not strictly from the dialogue.

2.5 Evaluation Metrics

The Exact match and F1 are used to evaluate the performance of models.

Exact match (EM): the number of answers predicted by a model and exactly matched the gold answers divides to the total number of gold answers in test set.

F1: F1 is computed by precision(p) and recall(r), where $p = N_{matched}/N_{predAns}$ and $r = N_{matched}/N_{goldAns}$. For a predicted answer, the precision(p) equals to the number of tokens that match to the gold answer($N_{matched}$) divided by the number of tokens in the predicted answer($N_{predAns}$), and the recall equals to the number of tokens that match to the gold answer divided by the number of tokens in the gold answer($N_{goldAns}$). For the Chinese dialogue data, to avoid the error of automatic Chinese word segmentation, the F1 score is calculated in the character level. Note that punctuation is ignored in calculating the EM and F1 scores.

3 Models

We evaluate representative neural end-to-end models for response geenration, which share a similar backbone of encoder-decoder structure (with the exception of DialoGPT, which has only a decoder). Below we give the common structure (Section 3.1) and then introduce the characteristic of each model (Section 3.2).

3.1 Model Structure Overview

We first give an overview of the structures of representative dialogue models. Most existing models adopt an encoder-decoder structure. As shown in Figure 1, the models consist of an utterance encoder, a context encoder and a decoder. Given a

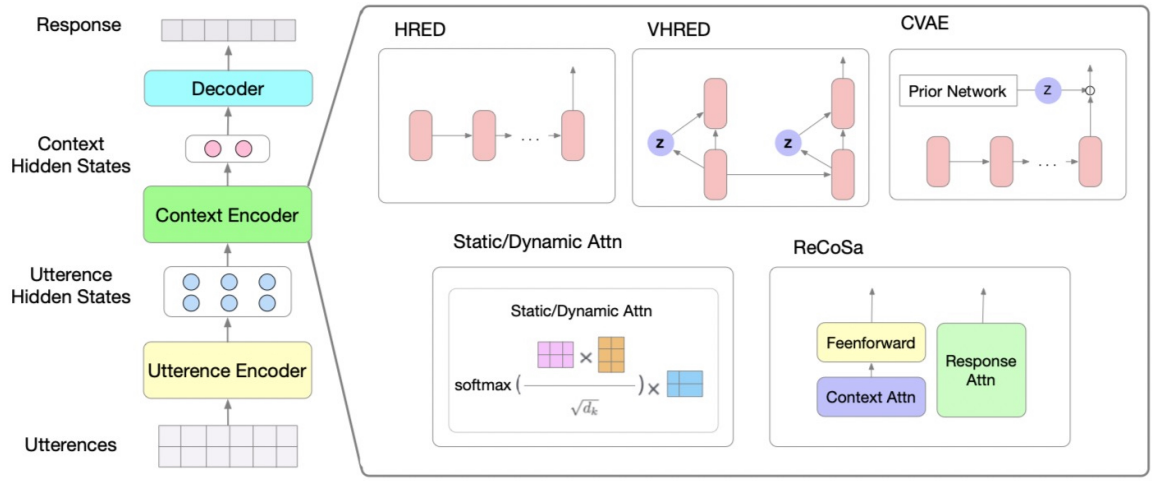

Figure 1: Model structure overview.

dialogue state, all the utterances in the current state, including past QA pairs, and a clarification request (question) are encoded as one input. The target is to generate the "gold" answer a_j from the input. The formalization of the process is:

$$y = \arg\max p(a_j|u_1, u_2, ..., q_i, \hat{a}_i, ..., u_k, q_j)$$

where $u_1, u_2, ..., u_k$ are dialogue utterance. q_i and $\hat{a}_i$ denote the i-th question and the predicted answer, respectively.

3.2 Representative Models

We choose the following 8 multi-turn dialogue generation models.

HRED: (Serban et al., 2016) is a hierarchical RNN-based encoder-decoder framework to sequentially model multi-turn dialogue and generate responses. It consists of two directional RNNs. One RNN is modeling the tokens in an utterance. The other RNN is modeling the utterances in a dialogue context.

vHRED: (Serban et al., 2017) is proposed to alleviate the generation of vague and generic response, which is caused by the gradient vanishing of HRED model, by introducing a hidden variable z. Therefore, vHRED is a variational enhanced HRED model.

CVAE: (Zhao et al., 2017) uses a prior network to model the gold response into a hidden variable z, which is as a condition in training step to improve the generation diversity.

Static/Dynamic Attention: the mechanisms (Zhang et al., 2018) alternatively model the contextual representations of multi-turn dialogue history using two types of attentions rather than using RNN.

ReCoSa: (Zhang et al., 2019) models the dialogue history in various granularity, e.g. context and response, using interactive attention and self-attention, respectively.

Transformer: (Vaswani et al., 2017) is used as a representative pretrained encoder-decoder model for dialogue generation.

DialoGPT: DialoGPT (Zhang et al., 2020) is a generative pretrained Transformer decoder for dialogue generation. To further conclude the characteristics of these models, Table 6 presents an overview of the characteristics of the chosen representative dialogue generation models in the proposed dialogue understanding task.

3.3 Implementation Details

For the training of the HRED, vHRED, CVAE and ReCoSa models, we use a bidirectional GRU (Cho et al., 2014) to encode the dialogue context and the input message. For the training of the static and dynamic attention models (Zhang et al., 2018), to be consistent to the setting in the original paper, a unidirectional GRU is utilized for contextual encoding of dialogue history. A fixed size of contextual window of dialogue utterances is used for modeling

	HRED	vHRED	CVAE	Static	Dynamic	ReCoSa	Transformer	DialoGPT
RNN	✓	✓	✓	✓	✓	✓	×	×
Attention	×	×	×	✓	✓	✓	✓	✓
Self-Attention	×	×	×	×	×	✓	✓	✓
Hidden variable	×	✓	✓	×	×	×	×	×
Encoder	✓	✓	✓	✓	✓	✓	✓	×
Decoder	✓	✓	✓	✓	✓	✓	✓	✓

Table 6: Characteristics of the representative dialogue models.

Transformer+Static	Transformer+Dynamic
$e_i = \frac{qk_i^T}{\sqrt{d_k}}$	$e_{i,t} = \frac{q_t k_i^T}{\sqrt{d_k}}$
$\alpha_i = \frac{\exp e_i}{\sum_j \exp e_j}$	$\alpha_{i,t} = \frac{\exp e_{i,t}}{\sum_j \exp e_{j,t}}$
O_i= MultiHead(q, k_i, k_i)	$O_{i,t}$= MultiHead(q_t, k_i, k_i)
$c = \sum_i \alpha_i O_i$	$c_t = \sum_i \alpha_{i,t} O_{i,t}$

Table 7: Implementation details of integrating the static and dynamic attentions into Transformer-based dialogue model.

the dialogue history[4].

For adding the Static attention into the Transformer model, the query q is the representation of a question. For the integration of Dynamic attention into the Transformer model, the query q_t denotes the decoded answer fragment in time step t. The key k_i denotes the output of encoding the i-th utterance in the dialogue context. The detailed modeling process is shown in Table 7. Please refer (Vaswani et al., 2017) for the definitions of q,k and d.

The dimension of the word and character embedding, which are initialized with GloVe[5], equals to 300. The RNN model is implemented with GRU. The size of hidden variable in vHRED and CVAE models is 300. For the ReCoSa model, the number of attention head equals to 6 and the number of self-attention layers is 3. Dropout is used in all models. For the experiments on English dialogue data, we use the 840B version of GloVe embedding. In experiments of Chinese dialogue data, to avoid the impact of different Chinese word segmentation tools, we use the character-level GloVe embedding, which is trained on Chinese Weibo corpus (Shang et al., 2015). Noted that the character embedding is fixed in the training process of these dialogue generation models.

[4]Note that we also verified the performance of non-fixed window size (e.g., a sliding window) but the performance of the above models all decrease.

[5]https://nlp.stanford.edu/projects/glove/

4 Results

Table 8 shows the results of the representative models on the proposed DEQA dataset. Overall, the model performances are below 20% in EM and below 40% in F1, which shows that the task is challenging for existing dialogue models. The results are relatively low compared to the same English benchmark on response generation task(Feng et al., 2020), which are in the range of 0.594 to 0.728 in averaged greedy matching and 0.548 to 0.746 in frequency-based similarity. The averaged greedy matching and frequency-based similarity are used to evaluate the coherence and informativeness of a generated response, respectively (Feng et al., 2020).

4.1 Results of Different Models

1) Attention-based models such as Static, Dynamic and ReCoSa, outperform the HRED, vHRED and CVAE models in EM score in both English and Chinese dialogue data and F1 score in Chinese data. It shows that attention/self-attention from an output token to the dialogue history context can be useful for capturing co-reference information.

2) Comparing the results in Table 8, the pretraining models outperform the representative dialogue models on both English and Chinese dialogue data, which demonstrates the superiority of the pretraining scheme on dialogue understanding. The results are consistent with results on the Winograd Scheme challenge (Ruan et al., 2019; Sakaguchi et al., 2020), which demonstrate that pretraining can be useful for co-reference resolution.

3) The CVAE model gives the best F1 score in Chinese dialogue data. Comparing the results of HRED and vHRED, we find that the use of latent variable may not improve the performances on the proposed task. Comparing the results of vHRED and CVAE, we can speculate that the improvements of performance may be from the introducing of prior information in context encoder.

4) The integration of static attention into Transformer model can further improve the performance

Model	English		Chinese	
	EM	F1	EM	F1
HRED	6.048	13.642	3.762	8.161
vHRED	5.882	14.954	2.871	7.239
CVAE	4.636	17.094	5.509	16.893
Static	6.048	18.132	7.052	14.873
Dynamic	6.214	17.602	6.856	14.408
ReCoSa	6.016	13.390	6.485	14.604
Transformer	9.959	23.221	17.723	33.306
+Static	10.133	24.155	**21.980**	**40.024**
+Dynamic	8.375	23.153	15.941	31.492
DialoGPT	**16.560**	**37.140**	19.307	33.798

Table 8: Results of the representative dialogue models on the proposed task.

on Chinese dialogue data. In addition, comparing the results of "Transformer" and "Transformer+Static", it indicates that the fine-grained encoding process can further improve the performance of the Transformer-based dialogue model.

4.2 Results on Different Answer Types

To further understand the main challenges, we split the test set into four subsets according to the answer types. Table 9 and 10 show the results of the models in English and Chinese dialogue data, respectively. Overall, the performance trends of DialoGPT and Transformer with static attention are consistent to the results in Table 8, which show the strong capability of pretraining models on the proposed task. Comparing each type of answers, we find that the difficulty of generating the answers in types of entity, phrase, clause and fragment is increasing in both English and Chinese dialogue data. One common reason is that the generation quality declines with the increasing of text length (Liu et al., 2018; Tan et al., 2020)[6]. In addition, the F1 score in English dialogue data is not monotonically decreasing as EM score. It is because the average number of tokens in English entities is close to 1, which leads to a lower F1 score than that in English phrases. It also reveals the reason that the EM and F1 scores in Entity type are closer than that in Phrase type.

5 Related Work

Conversational QA Recent research on conversational question answering (**ConvQA**) had been driven by two challenges, namely CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018). Rather than understanding the meaning of a given passage/document through the form of conversational question answering, the proposed task focuses on measuring the capability of understanding the dialogue itself. Besides the two challenges, several conversational machine reading/comprehension datasets were proposed (Elgohary et al., 2018; Dinan et al., 2018; Huang et al., 2018; Saeidi et al., 2018). The most common characteristic of these datasets are that their questions are open-domain and sequentially (or contextually) related, which shows a recent recognition in the research community that understanding the semantics of a complete conversation, including historical question and answer contexts, is crucial for these tasks. Our work is similar in spirit, but concentrating on clarification requests.

Clarification Request in Dialogue Clarification requests (**CR**) in dialogue are mainly motivated by acoustic understanding and semantic understanding (Schlangen, 2004; Stoyanchev and Johnston, 2015). They are used mainly as a way to establish mutual knowledge or grounding in communication (Gabsdil, 2003; Rieser and Moore, 2005). Purver et al. (2003) proposed to classify the forms of clarification requests into 8 categories, including non-reprise clarifications, reprise sentences, reprise sluices, reprise fragments, gaps, gap fillers, conventional and other. Rodríguez and Schlangen (2004) further summarized the surface forms, intonations and functions of clarification requests in spoken dialogue systems. Ginzburg (2016) detailed the semantics of dialogue and the fundamental problems to tackle for the semantic analysis in dialogue. In their work, a clarification request is defined to be a core function for dialogue systems to maintain the coherence of a dialogue.

This line of work coincides with our motivation that asking questions for clarification is a natural way to help understanding the meaning and maintaining coherence in dialogues. Therefore, the abilities of generating clarification requests to users and correctly responding to such requests from users are crucial to dialogue systems. Different from the above work, we build the DEQA, a Dialogue Entity via Question Answering dataset and investigate computational models for measuring the ability of machines on understanding the semantics of a dialogue via question answering.

[6]The average numbers of tokens in the answer types of entity, phrase, clause and fragment are 1.22, 2.15, 9.07, 5.48 in English dialogue data and 1.95, 3.28, 6.86 and 5.77 characters in Chinese dialogue data.

English / Model	Entity		Phrase		Clause		Fragment	
	EM	F1	EM	F1	EM	F1	EM	F1
HRED	7.707	9.985	6.987	19.819	8.333	17.384	3.579	14.624
vHRED	7.900	9.857	5.677	20.379	8.333	20.460	3.579	17.944
CVAE	5.545	12.439	4.933	22.246	8.333	20.241	2.004	21.298
Static	7.514	11.258	8.297	26.243	8.333	19.136	3.132	21.931
Dynamic	7.514	11.534	8.734	23.584	8.333	26.906	3.356	21.334
ReCoSa	0.000	8.853	0.901	23.589	0.000	22.651	0.000	22.702
Transformer	14.451	21.672	10.917	29.690	8.333	18.110	4.251	21.845
+static	15.414	22.240	12.227	31.014	8.333	24.934	2.908	22.823
+dynamic	13.295	18.822	8.297	33.305	8.333	**27.531**	2.685	22.811
DialogGPT	**24.085**	**39.194**	**20.961**	**44.281**	0.000	24.373	**6.040**	**31.480**

Table 9: Results of different answer types in English dialogue data.

Chinese / Model	Entity		Phrase		Clause		Fragment	
	EM	F1	EM	F1	EM	F1	EM	F1
HRED	5.609	9.658	2.970	7.638	1.282	4.684	0.000	4.931
vHRED	4.062	8.524	1.980	6.883	0.000	4.075	0.000	5.512
CVAE	8.952	19.860	3.902	15.747	1.282	13.019	0.658	11.575
Static	11.069	18.438	2.765	12.103	3.846	8.152	2.631	10.154
Dynamic	11.257	16.510	1.382	11.138	3.846	13.041	2.631	12.810
ReCoSa	2.421	4.945	0.980	4.685	0.000	3.726	0.000	4.064
Transformer	26.095	36.928	14.634	33.508	5.128	25.523	1.974	21.619
+static	**32.381**	**44.447**	**15.122**	**37.753**	7.692	**34.885**	**6.579**	**29.846**
+dynamic	24.000	34.903	11.707	31.280	6.410	26.449	0.658	22.635
DialogGPT	31.048	38.554	10.244	34.399	**8.974**	27.660	1.974	25.654

Table 10: Results of different answer types in Chinese dialogue data.

6 Conclusion

We proposed a novel evaluation task for co-reference resolution in dialogue understanding with a new benchmark dataset, DEQA. By asking a dialogue model to identify entity-oriented linguistic structures in dialogue history context, it directly measures the quality of dialogue understanding through response generation. Empirical comparisons show that the chosen representative dialogue models face challenges on the proposed benchmark dataset and clause and fragment types of co-references are paticularly challenging even for pretrained models. We will release the dataset for research use and further annotate the dataset with the questions that are related to more complex linguistic structures in future work.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 62076081, No. 61772153 and No. 61936010) and Science and Technology Innovation 2030 Major Project of China (No. 2020AAA0108605).

Ethics Statement

Our research is only oriented to semantic tasks and does not rely on any user information. To protect the privacy, intellectual property rights, and the rights of annotators, we took the following operations:

Privacy As we used crawler technology to get some raw data from Douban. It should be noted that to ensure the privacy of users, we directly deleted all User IDs and pictures during the crawling process, and only retained the text of the conversation. Then, we manually verify that all data has not leaked privacy.

Intellectual property rights Redistributing Douban's data may violate intellectual property rights. So we are now applying to Douban for permission to redistribute the data. We will release the dataset as soon as possible after obtaining permission. Before that, we first release the data of Dailydialog on the Github page and indicate that we are applying for the permission of Douban data.

Reasonable Compensation Regarding salary, Chinese annotation is easier than English for our Chinese annotators. Each piece of Douban data is paid about 1 yuan(about 16 cents), and DailyDialogue annotators are paid about 2 yuan (about 31 cents) per piece of data. Then we estimated the hourly salary of the annotator. The Chinese dataset annotator who is new to the task can complete about 70 annotations in an hour on average, so their hourly salary is at least 70 yuan(about 11 dollars) per hour.

The English dataset annotator completes an average of about 40 annotations per hour. The hourly salary is 80 yuan (about 12.5 dollars)per hour.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A customer service chatbot for E-commerce websites. In *Proceedings of ACL 2017, System Demonstrations*, pages 97–102, Vancouver, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1083.

Shaoxiong Feng, Hongshen Chen, Kan Li, and Dawei Yin. 2020. Posterior-gan: Towards informative and coherent response generation with posterior generative adversarial network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7708–7715.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060.

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35.

Jonathan Ginzburg. 2016. The Semantics of dialogue. In *The Cambridge Handbook of Formal Semantics*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. *arXiv preprint arXiv:1810.06683*.

Fang Kong, Min Zhang, and Guodong Zhou. 2019. Chinese zero pronoun resolution: A chain-to-chain approach. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–21.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *ICLR*.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. 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 4539–4549.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Verena Rieser and Johanna D Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 239–246. Association for Computational Linguistics.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality qa-srl annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Yu-Ping Ruan, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei. 2019. Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *arXiv preprint arXiv:1904.09705*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. *arXiv preprint arXiv:1809.01494*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th Workshop of the ACL SIG on Discourse and Dialogue*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. 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)*, pages 1577–1586.

Svetlana Stoyanchev and Michael Johnston. 2015. Localized error detection for targeted clarification in a virtual assistant. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5241–5245. IEEE.

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P Xing, and Zhiting Hu. 2020. Progressive generation of long text. *arXiv preprint arXiv:2006.15720*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1309–1318.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Weinan Zhang, Yiming Cui, Yifa Wang, Qingfu Zhu, Lingzhi Li, Lianqiang Zhou, and Ting Liu. 2018. Context-sensitive generation of open-domain conversational responses. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2437–2447.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, (Just Accepted):1–62.