Abg-CoQA: Clarifying Ambiguity in Conversational Question Answering

Anonymous authors

Abstract

Effective communication is about the dissemination of properly worded meaningful ideas/messages that are comprehensible to both sender and receiver and which ultimately can attract the desired response or feedback. For machines to engage in a conversation, it is therefore essential to enable them to clarify ambiguity and achieve a common ground. We introduce Abg-CoQA, a novel dataset for clarifying ambiguity in Conversational Question Answering systems. Our dataset contains 9k questions with answers where 1k questions are ambiguous, obtained from 4k text passages from five diverse domains. For ambiguous questions, a clarification conversational turn is collected. We evaluate strong language generation models and conversational question answering models on Abg-CoQA. The best-performing system achieves a BLEU-1 score of 12.9% on generating clarification question, which is 27.9 points behind human performance (40.8%); and a F1 score of 40.1% on question answering after clarification, which is 35.1 points behind human performance (75.2%), indicating there is ample room for improvement.

1. Introduction

Ambiguity is an intrinsic characteristic of human conversations and is particularly challenging in natural language understanding. People naturally resolve ambiguities in conversation by asking context-dependent clarification questions [Clark and Brennan, 1991]. Although there has been a surge in datasets/tasks on conversational question answering [Choi et al., 2018, Reddy et al., 2019], few studies have explored ambiguity resolution and clarification.

In this paper, we introduce Abg-CoQA, a dataset for clarifying the **amb**i**g**uity in **Co**nversational **Q**uestion **A**nswering. In Abg-CoQA, a machine has to answer an ambiguous question after resolving the ambiguity through a clarification dialog. As Figure 1 shows, the model needs to first detect whether a question (Q_i) in a conversation is ambiguous or not; for ambiguous questions it needs to generate a clarification question (CQ) targeting the ambiguity; since there are in general several possibilities for answering CQ, it then needs to provide an answer (A_i) based on each possible clarification reply (R_i). We develop Abg-CoQA with three main goals in mind.

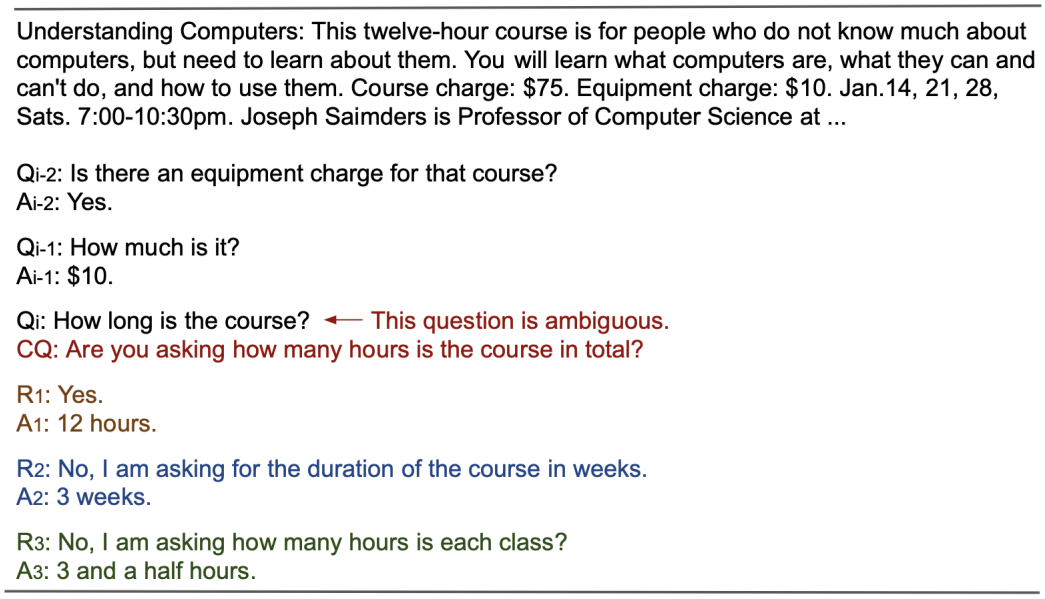

Figure 1: A conversation from the Abg-CoQA dataset.

The first concerns the nature of ambiguity in a human conversation related to a text passage. Our dataset covers the two main types of ambiguity in human conversations: when the question focus is ambiguous (*e.g.*, ambiguity in coreference resolution in Table 2); when there exist several possibilities to answer the question (*e.g.*, ambiguity in answer types in Table 3, Figure 1) [Ginzburg, 1996, Larsson, 2002]. The diversity of ambiguity types brings a challenge to models for generating appropriate clarification questions.

The second goal of Abg-CoQA is to ensure the naturalness of clarifying ambiguity in a conversation. In case of ambiguity in human conversations, the answerer asks a clarification question for collecting more information from the questionner. Different from a previous work on open-domain question [Min et al., 2020] which requires the model providing disambiguated rewrites of the ambiguous question, our dataset contains a natural clarification dialog for disambiguating the question.

The third goal of Abg-CoQA is to enable building question answering systems that perform robustly on the same question according to different clarification turns. The current CoQA datasets test the question answering systems' ability on understanding the passage and conversational history through answering a target question, which makes it hard to distinguish between a truly understanding of the context and a correct prediction based on superficial features [Chen et al., 2016, Weissenborn et al., 2017].

To summarize, Abg-CoQA has the following key characteristics: 1) it consists of 4k passages from five different domains and 9k conversational questions where 1k of them are ambiguous; 2) it covers four different ambiguity types and in most cases, the ambiguity is apparent after referring to the conversation history and researching all possible answers in the story; 3) each ambiguous question is followed by a clarification turn which consists of a question and several possible replies which lead to different answers to the originally ambiguous question.

We benchmark several deep neural network models, building on top of state-of-the-art conversational question answering and natural language generation models (Section 6). The best-performing system achieves an F1 score of 22.1% on predicting ambiguity, a BLEU-1 score of 12.9% on generating clarification question and a F1 score of 40.1% on clarification-based question answering. In contrast, humans achieve 40.8% BLEU-1 (27.9% higher) for clarification question generation and 75.2% F1 (35.1% higher) for clarified question answering, indicating that there is a lot of headroom for improvement.

2. Related Work

Conversational question answering requires a system to understand a text passage and answer a series of questions that appear in a conversation [Reddy et al., 2019, Choi et al., 2018]. It has potential applications on intelligent assistants and dialogue systems where ambiguity commonly exist. Clarifying ambiguity is essential for grounding in communication [Clark and Brennan, 1991]. To the best of our knowledge, all existing conversational question answering benchmarks assume each question has a single clear answer and ignore the possibility to be ambiguous. A recent work investigate the ambiguity in open-domain question answering [Min et al., 2020] and propose question rewriting for resolving ambiguity. However, question rewriting is not natural in conversation for clarifying ambiguity. Our work focuses on clarifying ambiguity in conversational question answering and resolves the ambiguity by interactively asking clarification questions, which follows the naturalness of communication [Traum, 1994, Kato et al., 2013].

Clarification questions have been used to resolve question ambiguity in other areas. Prior work studies the types, subjects and effectiveness of clarification questions that users ask on the Stack Exchange community question answering platform [Braslavski et al., 2017, Rao and Daumé III, 2018]. Our work differs from these two studies in that conversational question answering is in multi-turn and the clarification has a direct impact on the answer. Khalid et al. [2020] studies interactive communication with agent but focuses on communicative strategies for targeted, effective feedback about the system's understanding on reference tasks. Aliannejadi et al. [2019], Zamani et al. [2020] study a sequence of clarification questions to refine intents of simple searching query. Saeidi et al. [2018] worked on machine comprehension of natural language rules and considered clarification question for seeking missing information in the question during the communication with a robot assistant. Xu et al. [2019] define similar tasks on clarifying ambiguity as ours but their work is for knowledge-based question answering and the ambiguity types are only limited to entity reference and pronoun reference. In our work, the ambiguity naturally comes from human communication with related to a text passage, thus covers a diversity of sources (Table 2).

3. Task Definition

Figure 1 depicts the Abg-CoQA task. Given a passage P and a conversation $\{Q_{i-n}, A_{i-n},..., Q_{i-1}, A_{i-1}\}$ (where n is the number of the conversation turns), the task is to clarify the ambiguity in the next question Q_i if it is ambiguous. We consider three tasks.

Ambiguity Detection. Given a passage P and a conversation Q_{i-n}, A_{i-n},..., Q_{i-1}, A_{i-1}, detect whether the next question Q_i is ambiguous.

Clarification Question Generation. Given a passage P, a conversation $\{Q_{i-n}, A_{i-n},..., Q_{i-1}, A_{i-1}, Q_i\}$ where Q_i is ambiguous, generate a clarification question CQ which is helpful for disambiguating Q_i.

Clarification-based Question Answering. Given a passage P, a conversation $\{Q_{i-n}, A_{i-n},..., Q_{i-1}, A_{i-1}, Q_i, CQ, R_k\}$ where Q_i is an ambiguous question, CQ is a clarification question for Q_i, and R_k is one possible answer to CQ, answer the question Q_i which is no longer ambiguous based on the clarification. Note that there may exist several different answers to the clarification question CQ, and the answer to Q_i changes with the clarification answer.

4. Data Collection

We construct Abg-CoQA based on the CoQA dataset [Reddy et al., 2019]. Since most questions in the CoQA dataset are not ambiguous, we increase the ambiguity rate in our annotated corpus by 1) considering a partial conversation (keeping several previous conversational turns) rather than the full conversation; 2) pre-select probably ambiguous questions by using question answering models which are trained on CoQA dataset. We use Amazon Mechanical Turk (AMT) for crowdsourcing.

4.1 Collection Process

Given a story and a conversation (which is generally partial), annotators are asked to identify whether a question is ambiguous or not. If it is ambiguous, then provide a clarification turn. A clarification turn consists of a clarification question and all possible replies to it (could be one or several replies, see Figure 1). We also ask annotators to write all possible answers to the initial ambiguous question according to each clarification reply (refer to Appendix A for examples of the annotation interface).

In order to ensure the annotation quality on the crowd-sourcing platform, we filter AMT workers by location (US, CA, IN only for making sure workers are native English speaker), assignment approval rate (>97%) and customized qualification tests (for making sure workers understand the coding manual).

4.2 Ambiguous Question Pre-Selection

CoQA is a large-scale dataset for conversational question answering [Reddy et al., 2019]. We aim to annotate ambiguous questions in CoQA for our research purpose. According to their experimental results with a varied number of previous turns used as conversation history, all models succeed at leveraging history but the gains are little beyond one previous turn. They have same observation with human performance: given two history turns, human performance reaches up to almost same as given the full history. This suggests that most questions in a conversation have a limited dependency within a bound of two turns. Therefore, in our task, we only provide one or two history turns, which decreases the annotators' work load (shorter conversation) and potentially increases the ambiguity rate of questions (some questions may have longer dependency than 2 turns).

We pre-select questions which get a wrong answer given an incomplete history but could have been answered correctly given the full history. The intuition is that those questions

Domain	Total		Ambiguous		Abg rate
	#P	#Q	#P	#Q	%Q
Children's Sto.	296	636	71	90	14.2
Literature	991	2201	180	203	9.2
Mid/High Sch.	955	2172	186	226	10.4
News	909	1894	206	246	13.0
Wikipedia	817	1712	198	229	13.4
TOTAL	3968	8615	841	994	11.5

Table 1: Distribution of data numbers and ambiguous rates with respect to the domains in Abg-CoQA.

turn to be ambiguous because of the shorter conversation history rather than the inherent difficulty for answering the question itself. We first train a baseline model which has a BERT-based architecture with answer verification on the CoQA training dataset given the full conversational history as input. Then we select a sample if the model prediction given an incomplete history as input is greatly worse than given the full history. With this pre-selection process, we construct our corpus to be annotated.

Beside CoQA, we initially consider QuAC [Choi et al., 2018] as our data source since it is also a conversational question answering dataset based on context. We follow the same process for pre-selecting potentially ambiguous questions using a BERT-based model with history attention mechanism [Qu et al., 2019]. However, our pilot study on 50 samples of QuAC shows that the ambiguous rate is very low – 2%. We thus don't include QuAC in our work because of its low annotation efficiency.

With respect to the data splitting, we follow the same way as CoQA [Reddy et al., 2019]. For each source dataset (*e.g.*, Children's Story, Litterature, etc.), we split the data such that there are 100 passages in the development set, 100 passages in the test set, and the rest in the training set.

5. Data Analysis

The final dataset contains 3,968 passages and 8,615 questions, where 994 questions are annotated as ambiguous.

Domain Distribution. Table 1 shows the distribution of passages and questions with respect to the source domains of CoQA. We observe that the domain of Literature has the lowest ambiguous rate and the domain of Children's Story has the highest. This meets our intuition that language uses in Literature are more precise therefore there is less ambiguity in the conversation; in contradictory, Children's story is generally informal.

Types of Ambiguity. Table 2 shows a breakdown of the types of ambiguity in Abg-CoQA. We define a taxonomy with four categories, including ambiguity in coreference resolution, event references, time-dependency, and answer types. According to the two ambiguity types introduced in Ginzburg [1996], the ambiguity in coreference resolution is about the question focus and the ambiguity in answer types is about the answering possibilities; the ambiguity in event references and time-dependency cover the both ambiguity types. In comparison to Min et al. [2020], who studies ambiguity in open-domain questions, our

Type	Example
Coreference resolution (49%)	**Story**: ... Out of Africa(1985). Meryl is Karen Blixen, a Danish woman living in Kenya. The story follows Karen's attempts to run a coffee plantation and her love affair with ... $\mathbf{Q_{i-2}}$: What was her next movie? $\mathbf{A_{i-2}}$: Out of Africa. $\mathbf{Q_{i-1}}$: What type of character did she play? $\mathbf{A_{i-1}}$: Danish woman. $\mathbf{Q_i}$: What did she do? **Clarification question**: By "she" are you referring to Meryl Streep or her character? **Clarification reply #1**: I am referring to her character in Out of Africa. $\mathbf{A_i}$ **#1**: Her character, Karen, attempts to run a coffee plantation and has an affair. **Clarification reply #2**: I mean Meryl Streep. $\mathbf{A_i}$ **#2**: She is an actress who has worked in theatre, film, and television.
Time-dependency (23%)	**Story**: ... One of us grabbed a big wheel and rode it down the steep driveway into the street. Greg and I did it several times until the last time. The car hit him on the head. My brother and I both ran screaming just yelling for help and crying... $\mathbf{Q_{i-2}}$: Did anybody actually see the accident happen? $\mathbf{A_{i-2}}$: Yes. $\mathbf{Q_{i-1}}$: Who saw it? $\mathbf{A_{i-1}}$: My brother and I. $\mathbf{Q_i}$: What was everyone doing? **Clarification question**: Do you mean before the accident? **Clarification reply #1**: Yes. $\mathbf{A_i}$ **#1**: Riding a big wheel down the driveway into the street. **Clarification reply #2**: No, after the accident. $\mathbf{A_i}$ **#2**: We ran screaming and yelling for help and crying.
Answer types (16%)	**Story**: ... His ninth-grade English class for boys centers on books. "The novels they're reading now, are very manly novels. They're novels that deal with the arrogance of man and the pride of man." One of those books, for example, is "The Call of the Wild"... $\mathbf{Q_{i-2}}$: Who does he teach? $\mathbf{A_{i-2}}$: Boys. $\mathbf{Q_{i-1}}$: What are his pupils doing? $\mathbf{A_{i-1}}$: They're reading. $\mathbf{Q_i}$: What? **Clarification question**: The type of book or an example of a book they're reading? **Clarification reply #1**: The type of book. $\mathbf{A_i}$ **#1**: Novels that deal with the arrogance of man and the pride of man. **Clarification reply #2**: An example of a book. $\mathbf{A_i}$ **#2**: "The Call of the Wild".
Event references (12%)	**Story**: ... Dallas police named the suspected shooter, though CNN is not identifying him yet since he's a minor. The teen turns 18 in May, police said... $\mathbf{Q_{i-1}}$: How old? $\mathbf{A_{i-1}}$: 17. $\mathbf{Q_i}$: Was he identified by name? **Clarification question**: Do you mean identified by whom? **Clarification reply #1**: I mean by Dallas police. $\mathbf{A_i}$ **#1**: Yes. **Clarification reply #2**: I mean by CNN. $\mathbf{A_i}$ **#2**: No.

Table 2: Breakdown of the types of ambiguity in 50 random samples from ambiguous cases.

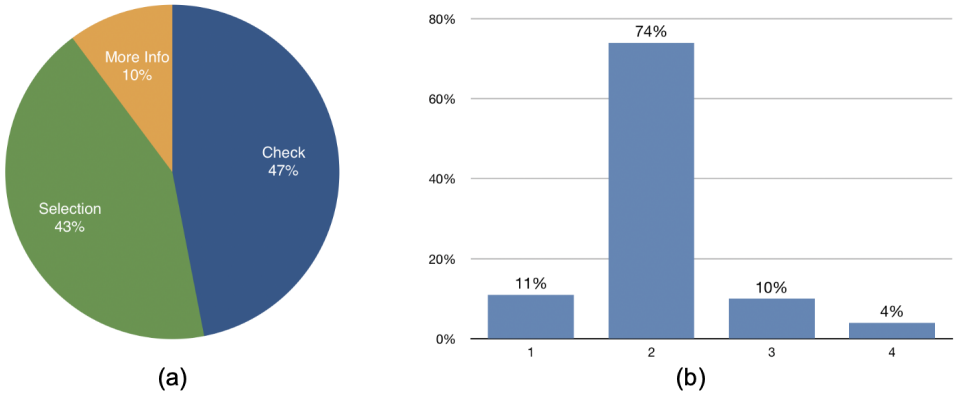

Figure 2: (a) Distribution of clarification strategies in 50 randomly sampled items from ambiguous cases. (b) Distribution of the number of clarification replies in all ambiguous items.

corpus contains one new ambiguity type – coreference resolution which is an inherent challenge in conversations. In addition, different from open-domain questions where more than a tier have the ambiguity in event references, it is actually a minor class in conversational questions since requested events are under the scope of the given story. In most cases, ambiguity is not apparent from the prompt question alone, but only after referring to the conversation history and researching all possible answers in the story.

Clarification Strategies. We classify the clarification questions in three types: More Information, Selection and Check. Our taxonomy follows Kato et al. [2013]'s work which classifies clarification requests of users in six categories, however, we only consider three types among them since we focus on the clarification strategy rather than the user intent. *Check* aims to confirm a hypothesis corresponding to the ambiguity (*e.g.*, the second example in Table 2); *Selection* aims to request an answer from two or more possibilities about the ambiguity (*e.g.*, the first and third example in Table 2); *More Information* directly asks for further details for clarifying the ambiguity (*e.g.*, the last example in Table 2). As shown in Figure 2(a), people prefer verifying their hypotheses (*e.g.*, *Check*, *Selection*) rather than asking open questions (*e.g.*, *More Info*) for clarifying the ambiguity.

With respect to the number of replies to the clarification question, we report the distribution in Figure 2(b). Most clarification questions have two different answers; 11% of them have only one reply; and 15% of ambiguous questions have more than two.

6. Models

To set initial performance levels on Abg-CoQA, we present a baseline model for each task. These tasks cover both the conversational question answering and language generation.

Ambiguity Detection. We formulate this task as the traditional question answering task by adding "ambiguous" as a possible prediction output. We consider two extraction-based models which have shown promising results for generating conversational responses on

the CoQA dataset as the baseline models for this task. Our baseline models are respectively build upon BERT [Devlin et al., 2019] and XLNET [Yang et al., 2019] plus prediction heads for each answer type [1] (respectively called BERT-AnsType and XLNET-AnsType). In order to take the *ambiguity* of questions into consideration, we append the "ambiguous" token at the end of the input passage. Therefore, the input to the model is the passage appended by "ambiguous", the conversation history and the question, and the expected output is the specific "ambiguous" token when the question is ambiguous; or the original response to the question when it is not ambiguous.

Clarification Question Generation. We fine tune a strong model for text generation – BART [Lewis et al., 2020] on our corpus as the baseline model for generating clarification questions. Prior work on BART demonstrates its effectiveness when fine tuned for news summarization. We append the given conversation history and the current question to the text passage and feed it into BART. The expected output is the clarification question. Since the ambiguous samples is in a small amount (*i.e.*, 1k), we also consider adding an additional fine-tuning prior to this clarification question generation task. In order to make the model learn to generate conversational questions, we first fine tune BART for generating the next question given the conversation history on the CoQA dataset (excluding the test set of Abg-CoQA). Then we fine tune the model on ambiguous samples for generating clarification questions.

Clarification-based Question Answering. We formulate this task as the original conversational question answering task by considering the clarification turn as the previous conversation history. In this task, we append the clarification turn (*i.e.*, a clarification question and one possible reply) to the passage and the conversation as the input sequence to the model. The expected output is the answer to the originally ambiguous question based on the clarification. We consider three different types of models as our baseline: the BERT-based model with answer verification (BERT+AnsType) which is also used for previous tasks, the XLNET model [Yang et al., 2019] with answer type prediction (XLNET+AnsType) which is a more powerful language model than BERT on question answering, and a generative model GPT-2 [Radford et al.] for zero-shot prediction. For BERT+AnsType and XLNET+AnsTyp, we first pre-train them on CoQA then fine-tune on Abg-CoQA by adding a clarification turn into the conversation.

7. Evaluation

7.1 Evaluation Metric

For answer generation, we use the same metric as CoQA: macro-average F1 score of word overlap. For computing a model's performance, each individual prediction is compared against n human answers resulting in n F1 scores, the maximum of which is chosen as the prediction's F1. For each question, we average out F1 across these n sets, both for humans and models. We follow the same way as CoQA for fixing the bias when computing human performance. In our evaluation on clarification-based question answering, n is equal to 3.

1. Answer type could be yes/no/unknown/extraction. For most cases, the prediction answers can be extracted from the input passage by predicting the start and end positions). However, if the answer is "yes"/"no"/"unknown", then the specific token may not exist in the passage thus need additional prediction heads for them.

Model	Question Answering (F1)						Ambiguity Detection		
	Child.Sto.	Literat.	M/H Sch.	News	Wiki	TOTAL	Precision	Recall	F1
BERT+AnsType	30.3	36.0	31.4	30.3	28.9	31.4	19.0	26.6	22.1
XLNET+AnsType	41.8	43.3	52.7	40.8	48.3	45.5	30.0	19.5	23.6

Table 3: Results on ambiguity detection.

For detecting ambiguity, we compute the precision, recall and F1 score as the evaluation metric on the two-class classification. We also report the macro-average F1 score on answer generation ($n = 1$) since we formulate predicting "ambiguous" as extracting an answer span from the passage.

For clarification question generation, we use BLEU scores as the main metric with a gold standard set of two human annotations.

7.2 Inter-rater Agreement

For measuring the inter-rater agreement on whether a question is ambiguous or not, we compute the Cohen's Kappa score on 68 randomly selected samples which are annotated by two Amazon Turk workers. The Cohen's Kappa score on the ambiguity detection is equal to 0.26, which shows a fair agreement. The key reason is that the ambiguity is subjective and personalized. In order to make it more objective, we randomly select 100 ambiguous cases (based on the previous annotation) in development set and ask three annotators to write an answer to each ambiguous question. Then we compute the macro-average F1 score of word overlap as a way to measuring the human agreement on answering ambiguous question. The F1 score is 65.3%, which is 23.5 points behind 88.8% F1 reported by CoQA, which reveals that ambiguous questions identified by our annotators are indeed difficult for human to provide consistent answers.

For measuring the inter-rater agreement on clarification questions, we ask a second Amazon Turk worker for annotating the clarification turn on ambiguous samples of the development and test sets. The F1 score is 45.3% and the BLEU-4 score is 21.9% in the test set (more BLEU scores are shown in Table 4). It is not surprising since there are different clarification strategies (Figure 2(a)) and annotators may have their own preference.

For the clarification-based question answering, we collect three annotations for each sample of the test set. The macro-average F1 score is equal to 75.2% and it shows the human performance on this task. Our study shows that the clarification increases 10 points for the inter-rater agreement (measured by F1 score) on answering ambiguous questions. We didn't train workers to write answers in the same style (*e.g.*, concise, short answers), so it is normal that the F1 score is lower than one reported in CoQA.

7.3 Results and Discussion

We report the experimental results of baseline models on our defined three tasks. The detailed experimental setting is introduced in Appendix B.

Ambiguity Detection. Since we consider ambiguity detection in the question answering setting, we report both the performance on answer prediction and on ambiguity detection in Table 3. With respect to the performance on question answering, the model trained on Abg-CoQA acheives an F1 score equal to 31.4% for BERT-based and 45.5% for

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Human Performance	40.78	31.57	26.24	21.86
fine-tune BART on Abg-CoQA	12.93	7.98	5.29	2.68
fine-tune BART on CoQA + Abg-CoQA	12.69	7.33	4.64	2.40

Table 4: Results on generating clarification questions.

Model	Children's Sto.	Literature	Mid/High Sch.	News	Wikipedia	**TOTAL**
Human	74.6	73.0	76.2	76.5	74.7	75.2
XLNET+AnsType	29.6	32.2	44.4	47.8	42.2	40.1
BERT+AnsType	30.2	31.4	36.6	49.1	40.9	38.7
GPT-2 zero-shot	14.5	7.0	11.5	17.4	6.9	11.8

Table 5: F1 scores on answer prediction after the clarification turn.

XLNET-based. It is not surprising because the baseline model failed on these samples when trained on CoQA (samples were selected by the pre-selection process; original average F1 score on them is only 3.8%). The best ambiguity detection performance gets a F1 score of 23.6%, which reveals the challenging of this task. We observe that the XLNET+AnsType baseline has a higher precision while the BERT+AnsType model has a higher recall.

Clarification Question Generation. Results for the clarification question generation task is shown in Table 4. We find that pre-training on CoQA on the question generation task doesn't help improve the performance, even slightly worse than directly fine-tuning BART on Abg-CoQA. We see a great gap between model and human performance. The difficulty mainly comes from identifying the ambiguous point in the question, so that the system can correctly generate a clarification question targeting the ambiguity.

Clarification-based Question Answering. Results for the answer prediction after the clarification is shown in Table 5. The model built upon XLNET achieves the best performance, however, still 35.1% behind the human performance, which shows that our task brings a new challenge to the question answering community. Even though samples in Abg-CoQA were pre-selected by the BERT-based baseline model, we don't see a great difference of performance between BERT and XLNET on this task. It demonstrates that the task is not biased towards the pre-processing of the BERT-based model. We also run GPT-2 as a representative of generative models on zero-shot prediction. Its performance decreases more than 40 points comparing to the reported F1 score (55%) on CoQA [Radford et al.].

We conduct an error analysis and find that existing strong models on standard conversational question answering tasks actually can't correctly answer the question based on different clarification replies. For example, a question asks "what is the color of the book?" and the story mentions two books respectively in red and green, thus the question is ambiguous. A clarification question is asked "Do you mean the first book or the second?". Models always predict "green" no matter the clarification reply is "the first" or "the second". It reveals that current models may be saturated to the training distribution rather than truly understand the conversation.

8. Conclusions

We introduce Abg-CoQA, a novel dataset for clarifying ambiguity in Conversational Question Answering systems. Our empirical study shows that it is challenging to identify ambiguity in a information-seeking conversation and generate clarification question. We propose clarification-based question answering as a benchmark task for evaluating the robustness of existing conversational question answering systems. We compare the performance of various models on this task and conclude that more research in conversational modeling is needed even though the performance on certain existing datasets is saturated.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484, 2019.

Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. What do you mean exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 345–348, 2017.

Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1223. URL https://www.aclweb.org/anthology/P16-1223.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL https://www.aclweb.org/anthology/D18-1241.

Herbert H Clark and Susan E Brennan. Grounding in communication. 1991.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Jonathan Ginzburg. Interrogatives: Questions, facts and dialogue. *The handbook of contemporary semantic theory. Blackwell, Oxford*, pages 359–423, 1996.

Makoto P Kato, Ryen W White, Jaime Teevan, and Susan T Dumais. Clarifications and question specificity in synchronous social q&a. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 913–918. 2013.

Baber Khalid, Malihe Alikhani, and Matthew Stone. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4417–4428, 2020.

Staffan Larsson. *Issue-based dialogue management*. Citeseer, 2002.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://www.aclweb.org/anthology/2020.emnlp-main.466.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, 2018.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, March 2019. doi: 10.1162/tacl_a_00266. URL https://www.aclweb.org/anthology/Q19-1016.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, 2018.

David R Traum. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science, 1994.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1028. URL https://www.aclweb.org/anthology/K17-1028.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and SUN Xu. Asking clarification questions in knowledge-based question answering. 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 1618–1629, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763, 2019.

Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3189–3196, 2020.

Appendix A. Examples of Annotation Interface

Figure 3 and Figure 4 show examples of the interface for AMT annotators.

Appendix B. Experimental Setup

For the BERT-based model, we adapt the SogouMRCToolkit[2] to our dataset and use its setting: batch size of 6, number epoch of 10, warm-up proportion of 0.1. We use the uncased-base BERT model as the backbone. The models are optimized using Adamax, with a learning rate of 3e-5. For the XLNET-based model, we adapt the XLNET extension toolkit[3]. The batch size is 8, the number of training steps is 6000. The model is optimized using Adam, with a learning rate of 3e-5. For the GPT-2, we follow its reported setting on the zero-shot CoQA task [Reddy et al., 2019]: add "Q" before each conversational question and clarification question and "A" before each answer as well as the end of the input sequence. For fine-tuning the BART model, we use the Fairseq toolkit [Ott et al., 2019]. We use the pre-trained large model, with Adam optimizer and learning rate of 3e-05. The total number of training steps is 2000 and the number of warm-up steps is 50. For generative models, *i.e.*, BART and GPT-2, we only consider the generated first sentence for evaluation since the number of tokens in generated text is in general defined larger than the ground truth.

2. https://github.com/sogou/SogouMRCToolkit
3. https://github.com/stevezheng23/xlnet_extension_tf

Figure 3: An example of the annotation interface when annotator selects *Non-ambiguous*. The interface updates with the chosen options.

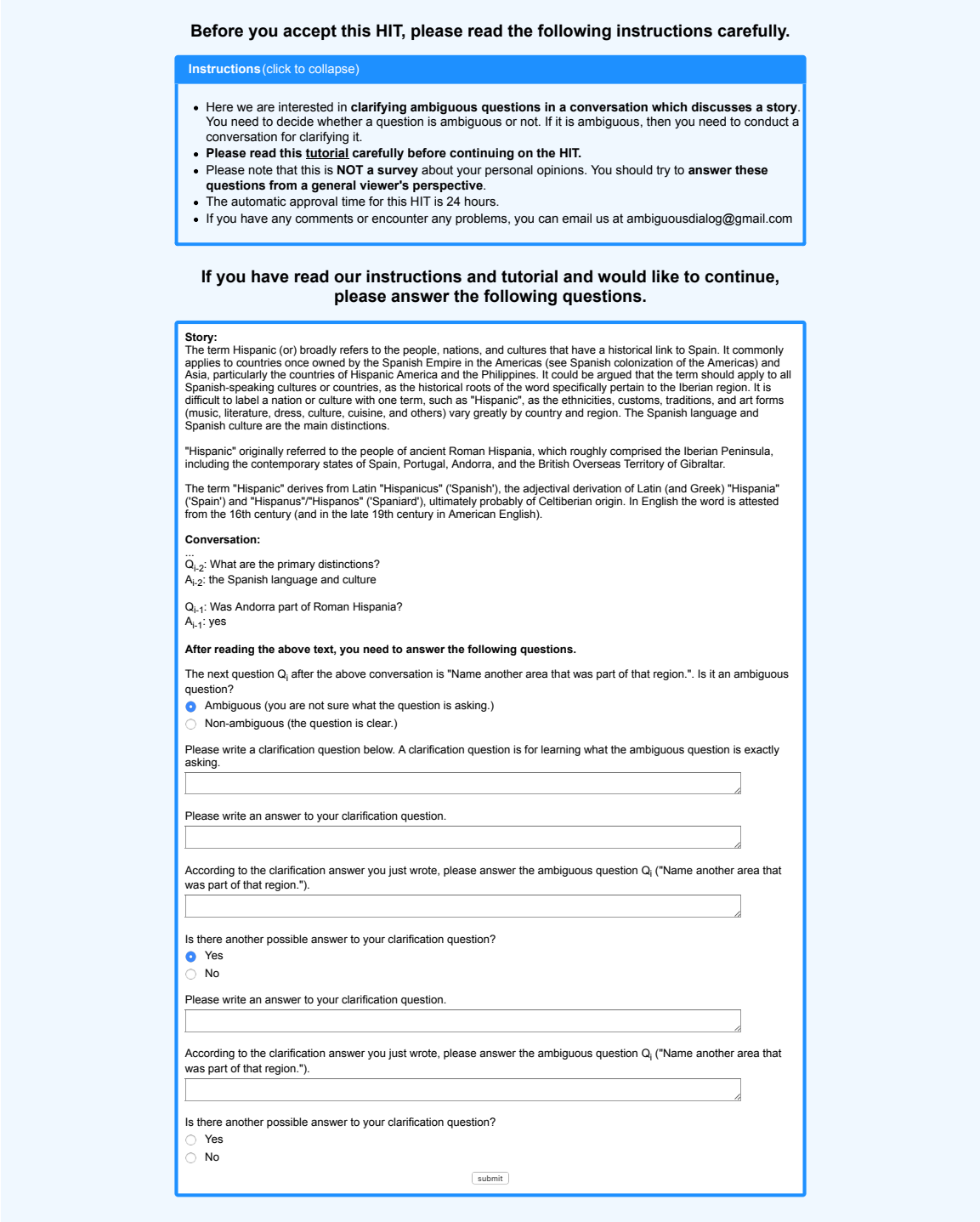

Figure 4: An example of the annotation interface when annotator selects *Ambiguous*. The interface updates with the chosen options.