Asking Clarification Questions to Handle Ambiguity in Open-Domain QA

Dongryeol Lee[1*], Segwang Kim[2*], Minwoo Lee[1], Hwanhee Lee[3], Joonsuk Park[4,5,6], Sang-Woo Lee[4,5,7], Kyomin Jung[1]

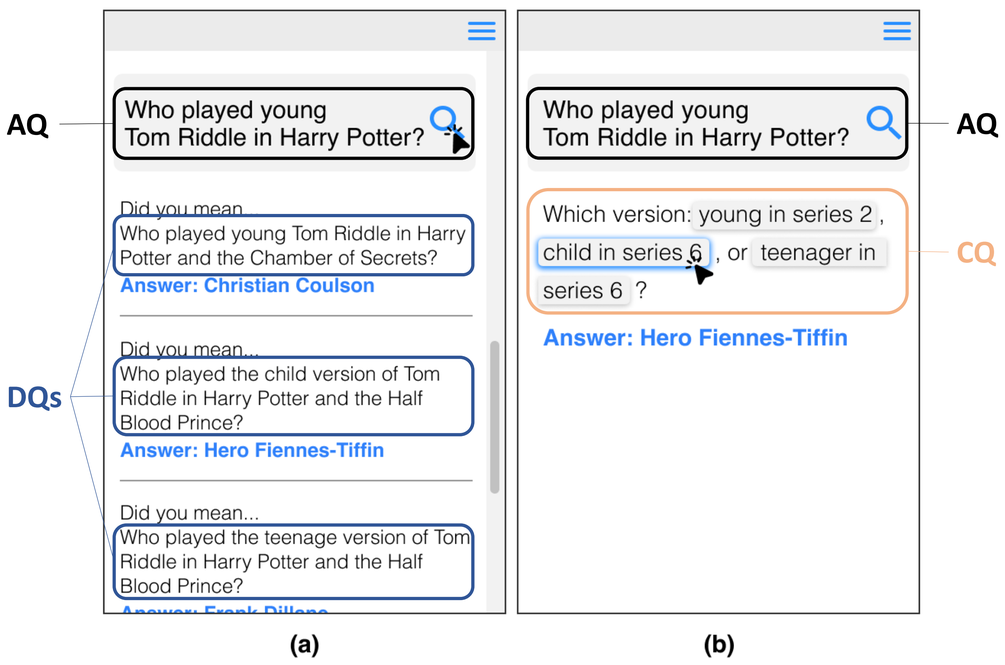
[1] Dept. of ECE, Seoul National University, [2] Samsung Electronic Mobile eXperience,
[3] Chung-Ang University, [4] NAVER AI Lab, [5] NAVER Cloud, [6] University of Richmond, [7] KAIST AI

Contributions

- **(a) We propose to use Clarification Questions (CQs) as a practical means to handle Ambiguous Questions (AQs) in Open-Domain Question Answering (ODQA).**
- **(b) We present CAMBIGNQ, a dataset to support CQ-based handling of AQs in ODQA.**
- **(c) We define a pipeline of tasks and appropriate evaluation metrics for CQ.**

Problem

- **Ambiguity** arises in Open-Domain Question Answering (ODQA) when there exist **multiple plausible answers** for the given **Ambiguous Question (AQ)**.



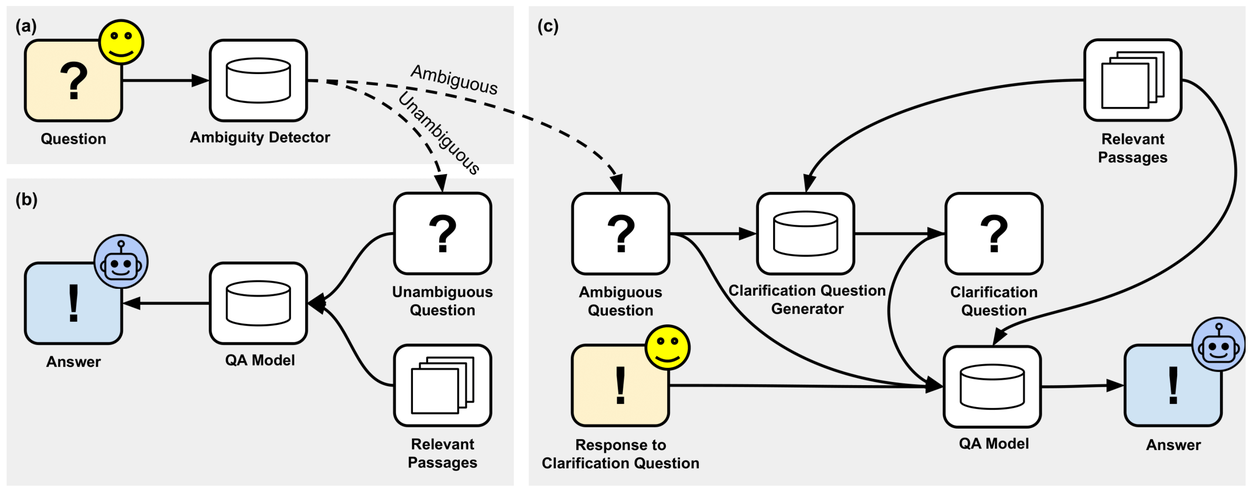
- **(a)** Previous works proposed **Disambiguated Questions (DQs)**, the minimally edited modification of AQ. They **enlist all DQs and corresponding answer**.
- **(b)** Instead, we propose **Clarification Questions (CQs)**, where the **user's response will help identify the interpretation** that **best aligns with the user's intention**.

Dataset: CAMBIGNQ

Split	CQ		Category	Options	
	#	Len.	Len.	Avg. #	Len.
Train	4,699	13.6	2.8	2.9	3.3
Validation	461	15.9	2.5	3.3	3.8
Test	493	17.8	2.8	3.4	4.1

- We collect high-quality of CQs by leveraging InstructGPT few-shot learning: **Generation via InstructGPT** and **Manual Inspection and Revision**.

Method



Our proposed method to handle AQs to in ODQA by asking CQs consists of three subtasks: **(1) Ambiguity Detection**, **(2) Clarification Questions generation**, and **(3) Clarification-based QA**.

(1) Ambiguity Detection

- Given a question q and relevant passages we classify whether q is ambiguous or not (binary classification)
- We test two settings: Direct Classification (**BERT**) and Generation-based Classification (**BART**).

(2) Clarification Questions Generation

- Given an AQ and relevant passages, we generate a CQ with following formats:

"$\boldsymbol{Which\ \underbrace{version}_{[category]}: \underbrace{young\ in\ series\ 2}_{[option_1]}, \underbrace{child\ in\ series\ 6}_{[option_2]}, or\ \underbrace{teenager\ in\ series\ 6}_{[option_3]}?}$"

- "$\boldsymbol{version}$" **is a** $[\boldsymbol{category}]$ to which all options belong.
- each $[\boldsymbol{option_i}]$ represent single interpretation of AQ.

(3) Clarification-based QA

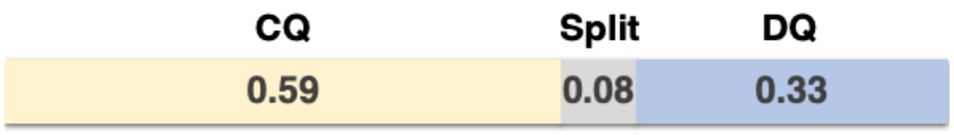
- Given an AQ, relevant passages, and a CQ, we generate a unique answer for every $[\boldsymbol{option_i}]$ by calling a QA model on AQ revised by CQ with following format:

"$\boldsymbol{Who\ played\ young\ Tom\ Riddle\ in\ Harry\ Potter.\ Which\ version: child\ in\ series\ 6?}$

Results

Human Preference Test (CQ vs DQ)

- Our proposed method **CQ (59%) is preferred over DQ (33%).**
- The prominent reasons for choice was its conciseness, ease of use, and clear guidance.



Ambiguity Detection Evaluation

- Direct Classification (No Answers for AQ) shows higher F1 compared to Generation based Classification (Predicted Answers for AQ)
- This is because the **average # of BART-generated answer is 1.24**, resulting in low recall.

Input in addition to AQ	Acc.	Pre.	Rec.	F1
No Answers for AQ	**63.9**	**61.9**	**60.7**	**61.3**
Predicted Answers for AQ	56.5	59.7	24.1	34.3

Clarification Questions Generation

- Evaluating generated CQs against gold CQs using automatic metrics (BLEU, BERTSCORE, EM) **can not capture semantic similarity.**

Input in addition to AQ and RPs	CQ		Category		Options			
	BLEU-4	BERTSCORE	EM	BLEU-1	Pre.	Rec.	F1	Avg. #
No Answers for AQ	**7.9**	**88.9**	20.2	**47.3**	**37.4**	18.2	24.5	2.0
Predicted Answers for AQ	**7.9**	**88.9**	**22.8**	44.0	36.9	**19.0**	**25.1**	2.0
Ground Truth Answers for AQ	15.4	89.6	25.2	46.9	34.3	34.4	34.3	3.7

Clarification-based QA

- The result shows **insufficient performance** across different settings. This is because the QA models, including LLMs, produce "**Same Answer**" **for the different questions**. Even though the QA model is prompted with Ground Truth CQ (Ideal case), the **model fails to capture subtle difference in the input**.

CQ used to clarify the AQ	NQ-pretrained BART				CQ-finetuned BART			
	Pre.	Rec.	F1	# Ans.	Pre.	Rec.	F1	# Ans.
CQ generated with No Answers for AQ	47.9	25.2	33.0	1.5	54.4	31.1	39.6	1.6
CQ generated with Predicted Answers for AQ	**49.6**	**26.2**	**34.3**	1.5	**55.4**	**32.0**	**40.5**	1.6
CQ generated with Ground Truth Answers for AQ	39.7	37.5	38.6	2.0	47.5	49.5	48.5	2.5
Ground Truth CQ	47.5	39.8	43.3	2.0	58.0	53.8	55.8	2.5

Code available at: https://github.com/DongryeolLee96/AskCQ