Covering Uncommon Ground: Gap-Focused Question Generation for Answer Assessment

Roni Rabin[1] **Alexandre Djerbetian**[1] **Roee Engelberg**[1,2] **Lidan Hackmon**[1]
Gal Elidan[1,3] **Reut Tsarfaty**[1,4] **Amir Globerson**[1,5]
[1] Google Research [2] Computer Science Dept., Technion
[3] Statistics Dept., Hebrew University of Jerusalem
[4] Computer Science Dept., Bar-Ilan University
[5] Blavatnik School of Computer Science, Tel Aviv University
{ronir, adjerbetian, roee, lidanh, elidan, reutt, amirg}@google.com

Abstract

Human communication often involves information gaps between the interlocutors. For example, in an educational dialogue, a student often provides an answer that is incomplete, and there is a gap between this answer and the perfect one expected by the teacher. Successful dialogue then hinges on the teacher asking about this gap in an effective manner, thus creating a rich and interactive educational experience. We focus on the problem of generating such gap-focused questions (GFQs) automatically. We define the task, highlight key desired aspects of a good GFQ, and propose a model that satisfies these. Finally, we provide an evaluation by human annotators of our generated questions compared against human generated ones, demonstrating competitive performance.

1 Introduction

Natural language dialogues are often driven by information gaps. Formally, these are gaps between the epistemic states of the interlocutors. Namely, one knows something that the other does not, and the conversation revolves around reducing this gap. An important example is the education setting where teachers ask students questions, and receive answers that may be incomplete. With the expectation of what a *complete* answer should contain, the teacher then engages in a gap-focused dialogue to help the student to arrive at a complete answer. There are multiple other application settings of information gaps, including support-line bots, long-form Q&A, and automated fact checking.

The core challenge in this setting is how to generate effective questions about the information gap. In terms of formal semantics and pragmatics, this gap can be viewed as the complementary of the *common-ground* (Stalnaker, 2002) held by the interlocutors. Somewhat surprisingly, despite much work on dialogue learning (Ni et al., 2022; Zhang et al., 2020) and question generation (Michael et al., 2018; Pyatkin et al., 2020, 2021; Ko et al., 2020), little attention has been given to generating questions that focus on such information gaps.

The formal traditional approach to representing the dialogic information gap is via the set of propositions that are known to one side but not the other (Stalnaker, 2002). However, this set can be quite large, and it is also unclear how to turn these propositions into dialogue utterances. We propose an arguably more natural representation: a generated set of natural language questions whose answers represent the information that the teacher needs to ask about to reduce the gap. We call these *gap-focused questions* (GFQs). A key advantage of this representation is that the generated questions can be used directly in the teacher-student dialogue.

Given a complete teacher answer and a partial student answer, there are many questions that could be asked, but some are more natural than others. For example, consider the complete answer *"A man is wearing a blue hat and a red shirt and is playing a guitar"*, and a student response *"There is a man playing the guitar"*. Two candidate questions could be *"What color hat is the man wearing?"* and *"What is the man wearing?"*. The second question is arguably more natural as it does not reveal information that is not in the teacher-student common ground, namely that a hat is being worn.

The above demonstrates some of the complexity of generating effective GFQs, and the need to rely on certain discourse desiderata. In this work we define the GFQ challenge, a novel question generation task, and we detail the desired properties of the generated questions. Subsequently, we provide a model for GFQ generation that aims to satisfy these desiderata, and demonstrate its competitiveness via a task of generating questions to fill the gap between premises and hypotheses in a standard *natural language inference* (NLI) setup.

In designing desired properties for GFQs, we take inspiration from theories of collaborative

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 215–227
July 9-14, 2023 ©2023 Association for Computational Linguistics

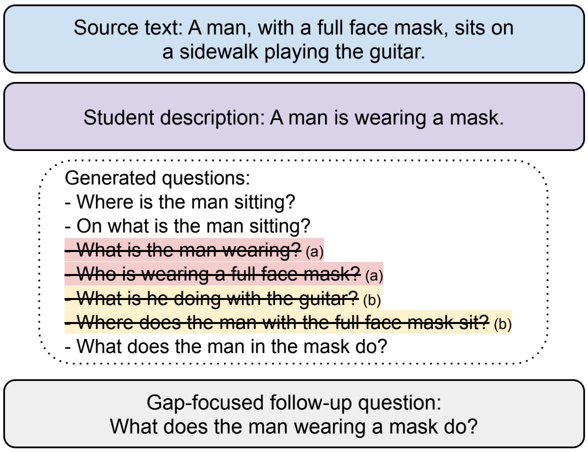

Source text: A man, with a full face mask, sits on a sidewalk playing the guitar.
Student description: A man is wearing a mask.
Generated questions:
- Where is the man sitting?
- On what is the man sitting?
- ~~What is the man wearing?~~ (a)
- ~~Who is wearing a full face mask?~~ (a)
- ~~What is he doing with the guitar?~~ (b)
- ~~Where does the man with the full face mask sit?~~ (b)
- What does the man in the mask do?
Gap-focused follow-up question: What does the man wearing a mask do?

Figure 1: Our Gap-Focused Question setup and approach. A student is asked to describe a source text from memory. The goal is to ask a follow-up question about information the student missed. Our approach is to generate a list of candidate questions, and then filter out ones that are either answerable from the student text (red strike-through (a)) or contain facts unknown to the student (yellow strike-through (b)). The follow-up question can be any of the remaining questions.

communication, and in particular Grice's maxims (Grice, 1975). For example, the *maxim of quantity* states that speakers are economic and do not communicate what is already known. Thus, the teacher should not ask about what is already in the common ground with the student. In the above example, this means not asking *"What is the man playing?"*. We describe additional desiderata in §3.

To tackle the GFQ challenge, we show how general-purpose NLP models (question generation, question answering, and constituency parsing) can be used to generate GFQs that satisfy the discourse desiderata. See Figure 1 for an outline of the process. To assess our model, we consider pairs of texts that contain information gaps, and evaluate our ability to capture these gaps using GFQs. Such texts are readily available in NLI datasets that contain pairs of a premise and an entailed hypothesis with less information. We consider the SNLI dataset (Bowman et al., 2015), and use human annotators to evaluate the merit of our approach relative to GFQs generated by humans.

Our contribution is three-fold. First, we propose the novel setup of gap-focused questions, a key element of a student-teacher discourse as well as other settings such as automated fact checking. Second, we identify desiderata inspired by conversational maxims, and provide a model for generating questions that satisfy them. Third, we demonstrate the merit of our model on an NLI dataset.

2 Related work

Natural dialogue is a key goal of modern NLP and, despite substantial progress, there is still a considerable difference between humans and models. In this work we focus on dialogues where the bot (teacher) knows more than the user (student), and the goal is to gradually decrease this knowledge gap via gap-focused follow-up questions.

Several works have focused on the problem of follow-up question generation in dialogues. However, to the best of our knowledge, none of these focus on information gaps as we do. Ko et al. (2020) introduce the problem of inquisitive question generation, where the goal is to generate questions about facts that are not in the text. This is not done in reference to a complete text, and is thus principally different from our goal. In fact, in our settings, an inquisitive question would typically be a bad GFQ, since it refers to information that is outside the knowledge of both teacher and student. Prior works considered a related task referred to as answer-agnostic question generation (Scialom et al., 2019), but with a focus on factual questions, whereas the inquistive setting is broader.

Another class of follow-up questions are clarification ones (Rao and Daumé III, 2018), which can also be viewed as a special case of inquistive questions. Again, there is no reference to a complete text that defines the information gap. Finally, there are works on follow-up questions guided by rules as in the SHARC dataset (Saeidi et al., 2018).

Our GFQ setting is also related to the challenge of explainable NLI (Kalouli et al., 2020), namely the task of explaining why a certain sentence entails another. The GFQ output can be viewed as a novel explanation mechanism of why the student text is entailed by the source text, as it explicitly refers to the gap between these texts.

Our work is inspired by novel uses of question generation models, particularly in the context of evaluating model consistency (Honovich et al., 2021). In these, question generation is used to find "LLM hallucinations" where the generated text is not grounded in a given reference text. Our task can be viewed as the inverse of the knowledge grounding task, and our particular focus is on the questions generated rather than just pointing to information gaps. An additional line of work in this vein is QA-based semantics, where text semantics are represented via a set of questions rather than a formal graph (e.g., see Michael et al., 2018).

3 Criteria for Gap-Focused Questions

Given a complete source text T_C and a student text T_S, our goal is to construct a model that takes T_S and T_C as input and produces a set of one or more questions Q that ask about the information gap between T_C and T_S. If one takes the term "information gap" literally, there are many such possible questions (e.g., which word appears in T_C but not in T_S). In a natural language setting we are obviously interested in questions that are *natural*, that is, would likely be asked by a human who knows T_C and has heard the student description T_S. When defining the desiderata for the generated questions, we consider what knowledge is held by the teacher and the student and what information is inside and outside their common ground (see Figure 2). We next identify desired properties for the generated questions, followed by a description of our model for generating gap-focused questions that satisfy these desiderata.

The following desired properties of an effective GFQ are loosely based on collaborative communication concepts (Grice, 1975):

- **P1: Answerability:** Only ask questions that can be answered based on the complete text T_C (areas $A \cup B$ in Figure 2). This follows from Grice's *maxim of relevance*; speakers say things that are pertinent to the discussion.

- **P2: Answers should not be in the common ground:** If the student has already demonstrated knowing a fact in T_S, there is no reason to ask about it again. Namely, in Figure 2, we don't want to ask about information in B. This pertains to Grice's *maxim of quantity*; speakers are economic, they do not utter information beyond the bare minimum that is necessary to ask the question, and they will refrain from repeating already-known information.

- **P3: Questions should only use information known to the user:** The question itself should rely only on information in T_S and not in T_C. For example if T_C is *"A Woman is wearing a blue hat"* and T_S is *"A woman is wearing something"*, it is preferable not to ask *"What color is the hat?"* as it refers to information that did not appear in T_S (i.e., that the woman is wearing a hat). This is loosely related to the Grice maxim of manner, where one tries to be clear, brief, and orderly. If we were to ask questions using information unknown to the user (in area A in figure 2), we may introduce unnecessary details and obscurity into the discussion.[1]

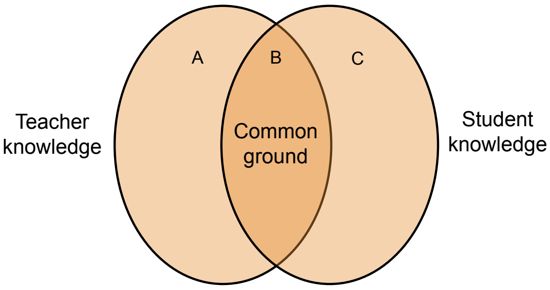


Figure 2: In our setup we consider the gaps between the teacher's knowledge (represented by the complete source text T_c, areas $A \cup B$ in the diagram) and the student's knowledge (represented by the student text T_s, areas $B \cup C$ in the diagram). We consider the information overlap between these two texts as the common ground between the teacher and the student (area B), which is a key component in defining good GFQs.

4 The GFQs Generation approach

We next describe our modeling approach for the GFQ generation problem, with the goal of capturing the properties described above. Before describing our GFQs generation approach, we briefly outline the NLP components we rely on in the question generation process:

A question generation model G that, given an input text T and a span $X \subset T$, generates questions about T whose answer is X.

A question answering model A, that takes as input a text T and a question Q about the text, and returns the answer or an indication that the question is unanswerable from the text.

A constituency parser P, that takes a text X, breaks it down into sub-phrases (constituents), and returns a parse tree.

Additional details about these components can be found in appendix C.

We are now ready to describe our approach for generating GFQs. The model generates an ordered set of possible follow-up questions Q_G via the following steps, which roughly correspond to the desired criteria described in §3:

Step 1: Generate answerable questions (P1). Using the constituency parser P, we extract the

[1] Note that in some cases this may only be partially possible and a "hint" must be provided in order to be able to phrase a grammatically correct and semantically sensible question.

spans of all the constituents in the source text T_C, except for those spanning the entire sentence, and single word spans containing functional elements (e.g., prepositions). For each span $X \subset T_C$, we use the question generation model G to generate a set of questions whose answer should be X, thus creating a set of questions that satisfy the answerablity property. We denote this set Q_T and assign $Q_G = Q_T$.

Step 2: Filter questions whose answers are in the common ground. (P2). We next wish to remove questions that are answerable by the student text T_S. To that end, we use the question answering model A, and for each $q \in Q_G$ if $A(T_S, q) \neq$ "UNANSWERABLE", we set $Q_G = Q_G \setminus \{q\}$.[2]

Step 3: Prefer questions which only use information known to the user (P3). We prefer questions that do not reveal information beyond what is known to the user. This is not always strictly possible and thus, instead of filtering, we rank questions according to the (possibly zero) amount of additional information they reveal. To do so, let R be all the answers to the questions in Q_G. By construction R contains spans from T_C that the student didn't mention, i.e. these are spans that we would prefer not to appear in the generated questions. For each $q \in Q_G$, we count the number of items in R included in q. We sort Q_G in ascending order by this number and return the first element. We thus return a question that uses the least number of facts unknown to the student.

5 Experiments

We next describe an evaluation of our GFQ model.

Data: We use the *SNLI Dataset* (Bowman et al., 2015) where a Natural language inference (NLI) pair contains two sentences denoting a premise and a hypothesis, and the relation between them can be *entailment*, *contradiction* and *neutral*. We focus on pairs labeled as entailment, and filter out those with bi-directional entailment, so that there is a gap between hypothesis and premise. We do not use any data for training, and apply our model to the test partition of the SNLI dataset.

Evaluation Benchmark: In order to compare the quality of our automatically generated questions to manually generated ones, we asked human annotators to generate questions for 200 instances of the SNLI test set (see Appendix A for the annotator instructions). We emphasize that these questions were only used for evaluation, as explained below, and not for training the model. They were collected after model design was completed. We release this evaluation dataset to the public, it is available here. See additional details about this dataset in appendix E.

Model	Average score
Step 1	3.72
Step 2	3.86
Step 3	3.94
Human	4.06

Table 1: Average scores of the different generation methods on 200 questions, each rated by 3 annotators.

Annotator Evaluation of Generated Questions: As with other generative settings, offline evaluation is challenging. In fact, even if we had human generated questions for all SNLI, using those for evaluation would need to assume that they are exhaustive (otherwise the model can generate a good question but be penalized because it is not in the set generated by humans). Instead, as is commonly done (Ko et al., 2020), we rely on human evaluation. We present annotators with T_C, T_S and a candidate GFQ q and ask them to provide a $1-5$ score of how well q functions as a follow-up question (see Appendix A for annotators instructions). We use 3 annotators per question.

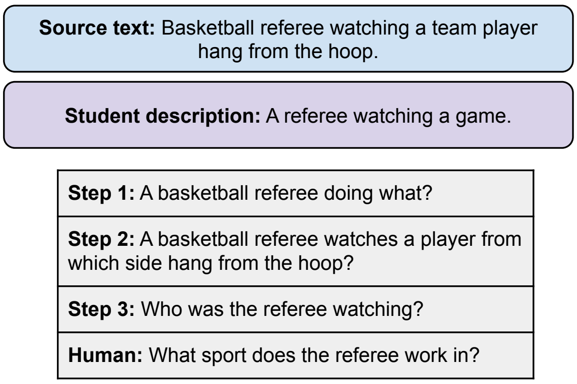
Compared Models: We compare four generation approaches: **Human**: Questions generated by human annotators; **Step 1**: This model selects a random question out of those generated by the question generation model (i.e., Step 1 in §4). We note that this is already a strong baseline because its questions are based on the source text. **Step 2**: The outcome of Step 2 in §4 where only questions not answerable by the student text are kept. **Step 3**: The outcome of Step 3, where we additionally aim for questions which use information known to the user.

Results: Table 1 provides the average scores for each of the considered models and the human generated questions. It can be seen that each step contributes to the score, and human generated questions are somewhat better than our final model

[2] Note that Step 2 will also filter out questions that the student answered incorrectly. This would be an area for improvement in future models.

Source text	Student description	Generated question (Step 3)
A man stands by two face structures on Easter Island.	A man on Easter Island.	Two faces are what on Easter Island?
Two young children, one wearing a red striped shirt, are looking in through the window while an adult in a pink shirt watches from behind.	A person in a shirt.	What is one child wearing?
A man in a purple jersey is falling down while chasing a player in a green jersey playing soccer	The two soccer players run around chasing each other	What is the man in the cartoon wearing?

Table 2: Examples of the loss patterns found in the analysis of low scoring questions. See details in the Error Analysis paragraph in section 5.



Source text: Basketball referee watching a team player hang from the hoop.

Student description: A referee watching a game.

Step 1: A basketball referee doing what?

Step 2: A basketball referee watches a player from which side hang from the hoop?

Step 3: Who was the referee watching?

Human: What sport does the referee work in?

Figure 3: An example of the steps of our Gap-Focused Questions model, and a human-generated question.

(**Step 3**). Using the Wilcoxon signed-rank test for paired differences, we found that all differences were significant at p-value ≤ 0.05.

Examples: Figure 3 shows an example of the three stages, and a human generated question. Appendix F provides more examples.

Error Analysis: We analyze cases where our final model (Step 3) received low scores from the annotators (an average score of 3 and lower). In our analysis we have observed three main loss patterns (sometimes appearing together): **(1)** Poor question phrasing — these are questions whose structure or choice of words is less natural than if a person were to ask the same question. See example in the first row in Table 2. **(2)** Questions which include information outside of the teacher-student common ground. These are cases where the minimum criterion defined in Step 3 still results in a question with some information unknown to the user. See examples in the first 2 rows in Table 2. **(3)** Questions including information outside the complete source text. In rare cases we have found that the question generation model generates questions that include "hallucinations" or point to issues in the semantic understanding of the complete source text. See the third example in Table 2.

6 Conclusion

We consider the task of question generation in a novel setting where there is an information gap between speakers, and the gap-focused questions (GFQs) aim to reduce this gap. Building on advances in question generation and question answering, we show how to generate useful GFQs that meet several natural criteria inspired by theories cooperative conversation. It is natural to ask whether one can employ a fully generative approach for GFQs using LLMs. This is a natural direction for future study, and we believe that the criteria and design choices we studied here will be significant in defining and evaluating such future work.

Limitations

We present the first study of generating questions for filling in information gaps. Our method is limited in several ways. First, it focuses on information that is explicitly missing, and does not discuss information that is inaccurate or incomplete in other ways. Second, it only asks one follow-up question and does not address multi-turn dialogue about a student answer, or multiple student answers. Finally, our approach makes somewhat restricted use of the student answer, and it will be better to generate questions that directly uptake information from the student text (Demszky et al., 2021). We leave the deep investigation of these for future work.

Acknowledgments

We thank Avi Caciularu for constructive feedback on this work.

Ethics and Impact

Regarding risks, as with any NLP model, care must be taken in application, so that it generates truthful information, and does not introduce biases. However, we think this is not a major concern in our case as our modeling will generate text directly related to the source and student texts. In terms of impact, our approach can be used to improve a wide array of applications, including educational dialogue (e.g., reading comprehension), supportline bots, and automated fact checking.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905—-3920.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Valeria de Paiva, Richard Crouch, and Mennatallah El-Assady. 2020. XplaiNLI: Explainable natural language inference through visual analytics. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 48–52, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke S. Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *NAACL-HLT*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1429–1441, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. 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.

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. In *EMNLP*.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 6027–6032.

Robert Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5/6):701–721.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

A Annotating Guidelines

Here we provide all the guidelines to annotators, for both human question generation and human rating of questions generated by the model.

Guidelines for the human annotator task of writing follow-up questions: We depict the guidelines and the examples for the writing follow-up questions task in Figure 4, and the task design in Figure 5.

Guidelines for the human annotator task of rating follow-up questions: We depict the guidelines of the task of rating the follow-up questions in Figure 6, the examples in Figure 7, and the task design in Figure 8.

B Annotator Related Information

Annotators were paid by the hour, and recruited as contractors for a variety of annotating projects by our team and related teams. The annotators are all native English speakers (from Canada and the US). They are also aware of the way in which the information will be used. There are no special ethical sensitivities in the collection process and thus it was exempt from an ethics review board.

C Implementation Details

Question Generation Model: As our question generation model G, we use the T5-xxl model (Raffel et al., 2020) fine-tuned on SQuAD1.1 (Rajpurkar et al., 2016). We also use beam search and question filtering, similarly to Honovich et al. (2021, Section 2), see this work for further details.

Question Answering Model: For our question answering model A, we use the T5-xxl model (Raffel et al., 2020) fine-tuned on SQuAD2.0 (Rajpurkar et al., 2018).

Constituency Parser: We use the Berkeley Neural Parser (Kitaev and Klein, 2018), implemented in the spaCy package.[3]

SNLI Filtering: We consider the subset of SNLI with an "entailed" label. Since we are not interested in the case of equivalent hypothesis and premise, we filter out bi-directional entailments using an NLI model (similar to (Honovich et al., 2022)). In the resulting set of one-directional entailments, the information in the premise (T_C) is *strictly* greater than the information in the hypothesis (T_S), which is our case of interest.

D Computational Resources Details

In terms of computational resources, the project is lightweight, as it required no training at all, and just running inference steps of pre-trained models (question answering, question generation and parsing), all of which run in several minutes on standard GPUs.

E GFQ test released dataset

We release a benchmarking dataset of 200 examples from SNLI test with a human generated gap-focused question. The data is available here.

Details about the dataset We asked 3 annotators to write questions for each SNLI pair (see guidelines in appendix A) and used a heuristic to select a single GFQ. When selecting this single question our goal is to prefer GFQs where multiple annotators chose to write a question about the same topic. We therefore apply the following heuristic: for each human written question q we used our question answering model A and define a as the answer to this question given T_c: $a = A(T_c, q)$. We then count n: the number of annotators which produced questions leading to the same answer a, we look at the questions for which n is maximal and choose a random question from there.

License This data as well as the underlying SNLI data are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License [4].

F Examples of Generated Questions

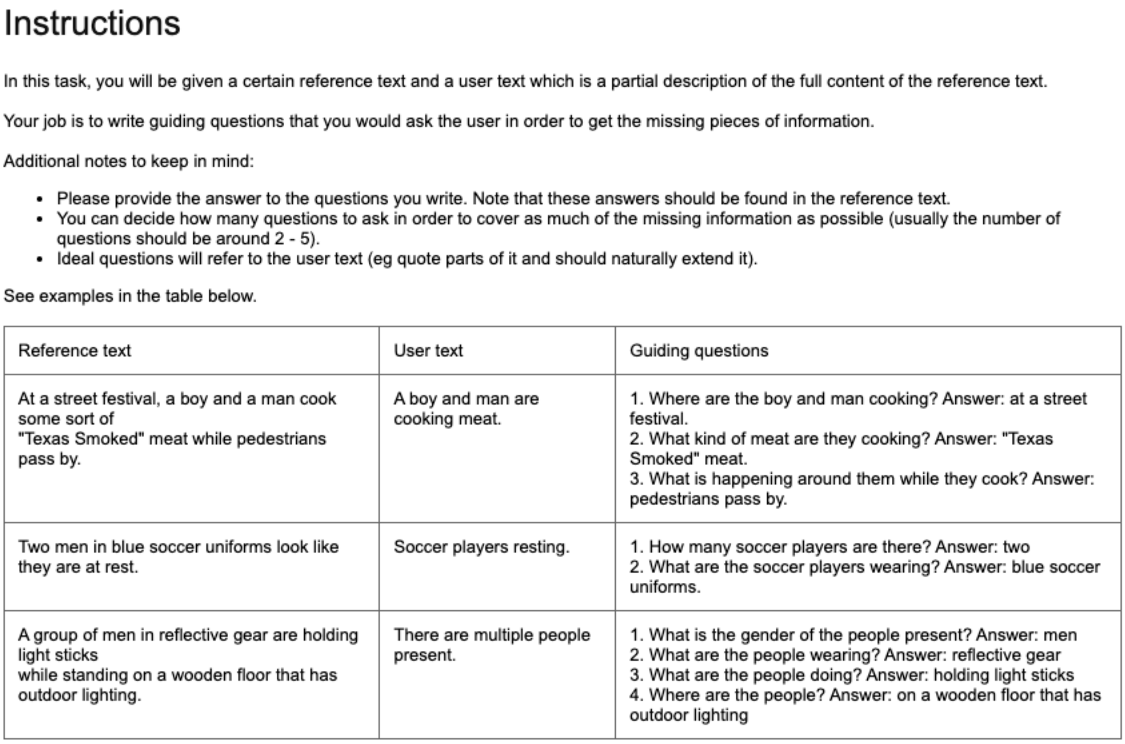
Here we provide examples of questions generated by humans and by the different models we consider. Table 3 reports questions generated by Step 1, Step 2, Step 3 and Human.

G Data Related Information

The data collected from annotators contains the manually generated questions and the scoring of generated questions. There are no issues of offensive content or privacy in this data, as it based closely on the SNLI dataset.

[3] We used spaCy3.0 – `https://spacy.io/`.

[4] http://creativecommons.org/licenses/by-sa/4.0/



Instructions

In this task, you will be given a certain reference text and a user text which is a partial description of the full content of the reference text.

Your job is to write guiding questions that you would ask the user in order to get the missing pieces of information.

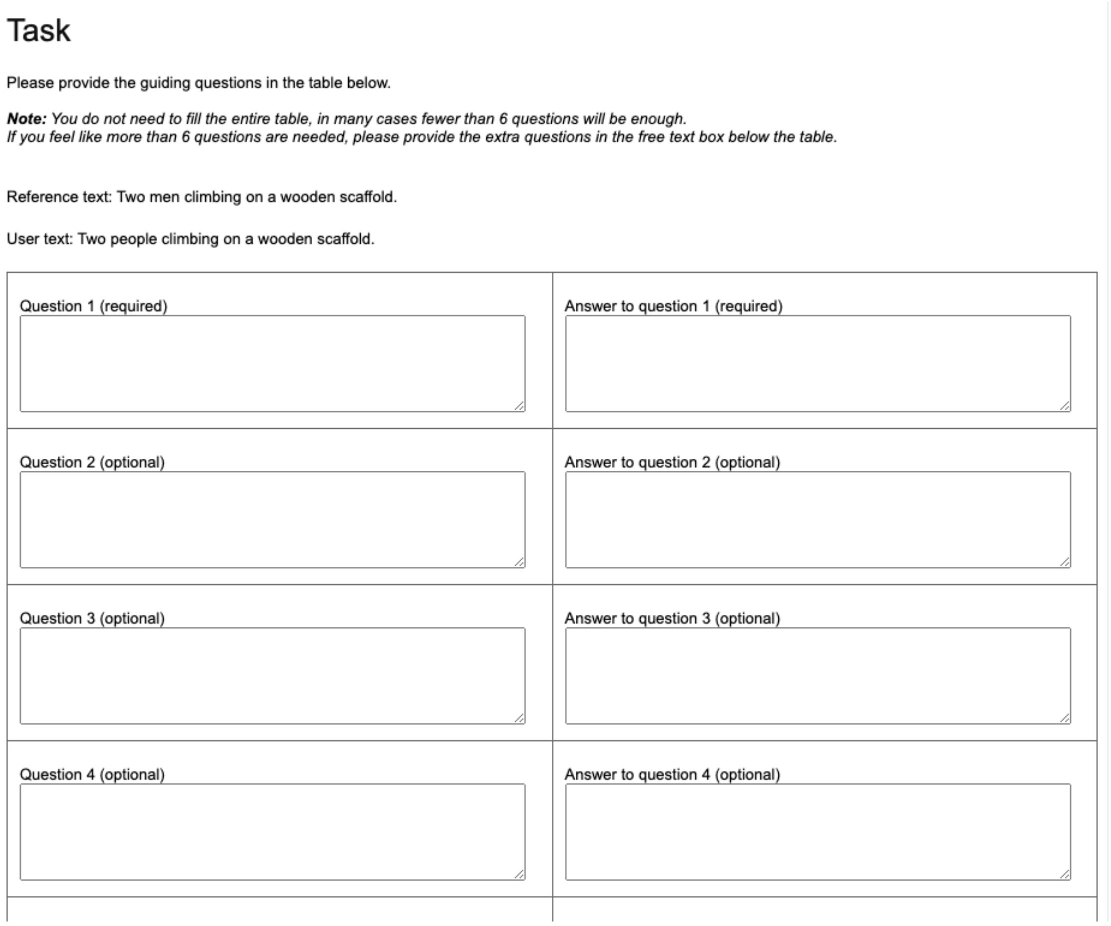
Additional notes to keep in mind:

- Please provide the answer to the questions you write. Note that these answers should be found in the reference text.
- You can decide how many questions to ask in order to cover as much of the missing information as possible (usually the number of questions should be around 2 - 5).
- Ideal questions will refer to the user text (eg quote parts of it and should naturally extend it).

See examples in the table below.

Reference text	User text	Guiding questions
At a street festival, a boy and a man cook some sort of "Texas Smoked" meat while pedestrians pass by.	A boy and man are cooking meat.	1. Where are the boy and man cooking? Answer: at a street festival. 2. What kind of meat are they cooking? Answer: "Texas Smoked" meat. 3. What is happening around them while they cook? Answer: pedestrians pass by.
Two men in blue soccer uniforms look like they are at rest.	Soccer players resting.	1. How many soccer players are there? Answer: two 2. What are the soccer players wearing? Answer: blue soccer uniforms.
A group of men in reflective gear are holding light sticks while standing on a wooden floor that has outdoor lighting.	There are multiple people present.	1. What is the gender of the people present? Answer: men 2. What are the people wearing? Answer: reflective gear 3. What are the people doing? Answer: holding light sticks 4. Where are the people? Answer: on a wooden floor that has outdoor lighting

Figure 4: Human annotator guidelines and examples for the task of writing follow-up questions.



Task

Please provide the guiding questions in the table below.

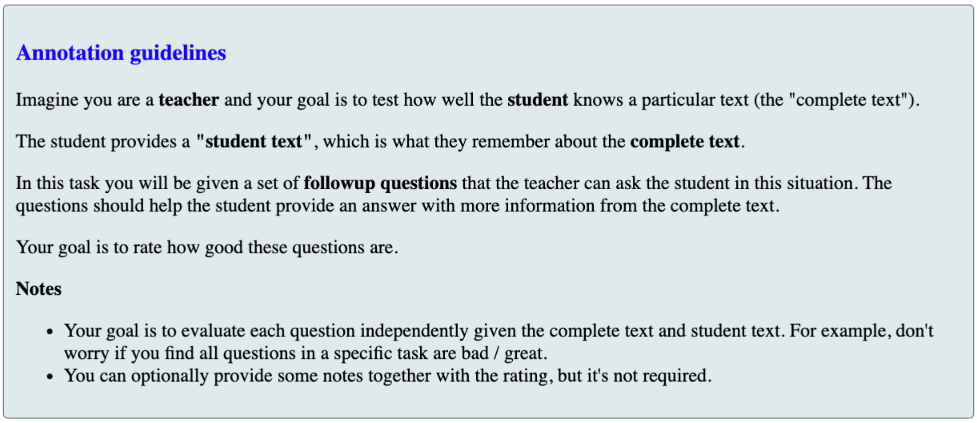
Note: *You do not need to fill the entire table, in many cases fewer than 6 questions will be enough.*
If you feel like more than 6 questions are needed, please provide the extra questions in the free text box below the table.

Reference text: Two men climbing on a wooden scaffold.

User text: Two people climbing on a wooden scaffold.

Question 1 (required)	Answer to question 1 (required)
Question 2 (optional)	Answer to question 2 (optional)
Question 3 (optional)	Answer to question 3 (optional)
Question 4 (optional)	Answer to question 4 (optional)

Figure 5: The user interface of the human annotator task of writing follow-up questions.



Annotation guidelines

Imagine you are a **teacher** and your goal is to test how well the **student** knows a particular text (the "complete text").

The student provides a **"student text"**, which is what they remember about the **complete text**.

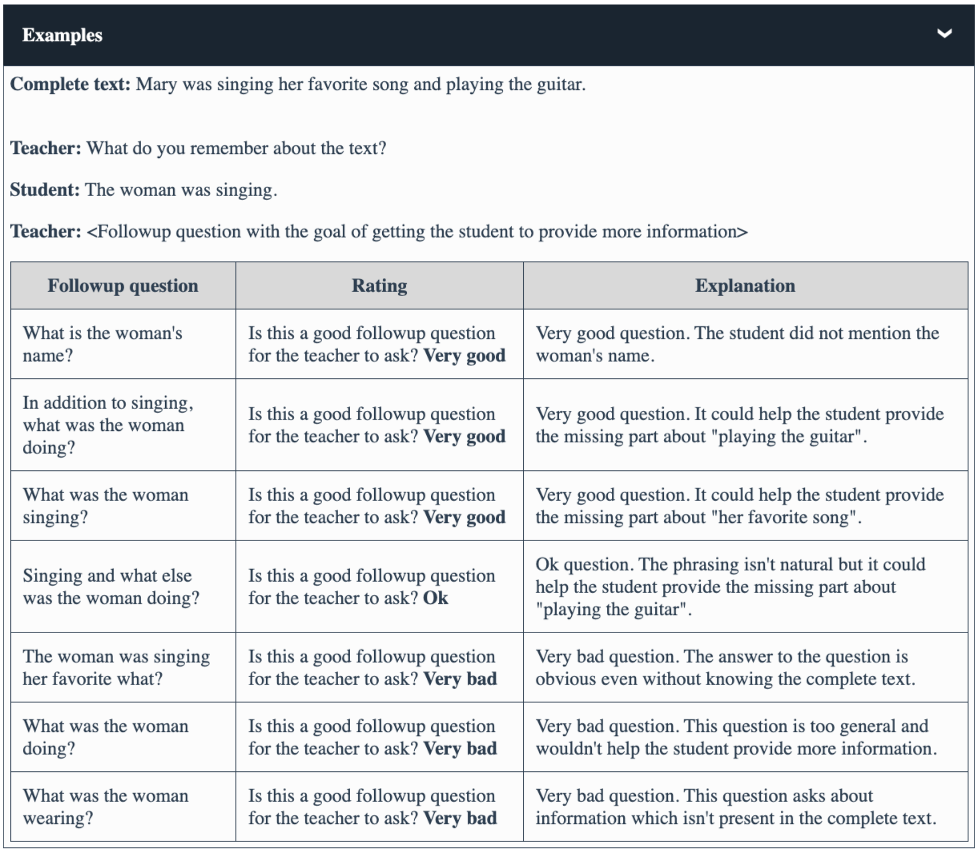
In this task you will be given a set of **followup questions** that the teacher can ask the student in this situation. The questions should help the student provide an answer with more information from the complete text.

Your goal is to rate how good these questions are.

Notes

- Your goal is to evaluate each question independently given the complete text and student text. For example, don't worry if you find all questions in a specific task are bad / great.
- You can optionally provide some notes together with the rating, but it's not required.

Figure 6: Guidelines for the human annotator task of rating follow-up questions.



Examples

Complete text: Mary was singing her favorite song and playing the guitar.

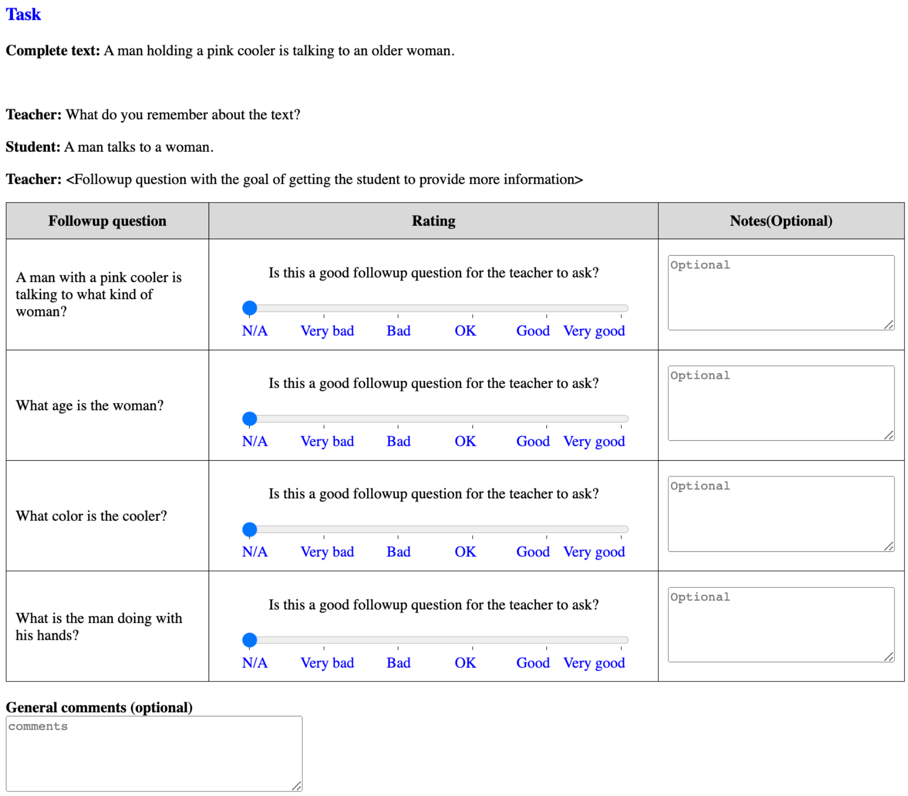
Teacher: What do you remember about the text?

Student: The woman was singing.

Teacher: <Followup question with the goal of getting the student to provide more information>

Followup question	Rating	Explanation
What is the woman's name?	Is this a good followup question for the teacher to ask? **Very good**	Very good question. The student did not mention the woman's name.
In addition to singing, what was the woman doing?	Is this a good followup question for the teacher to ask? **Very good**	Very good question. It could help the student provide the missing part about "playing the guitar".
What was the woman singing?	Is this a good followup question for the teacher to ask? **Very good**	Very good question. It could help the student provide the missing part about "her favorite song".
Singing and what else was the woman doing?	Is this a good followup question for the teacher to ask? **Ok**	Ok question. The phrasing isn't natural but it could help the student provide the missing part about "playing the guitar".
The woman was singing her favorite what?	Is this a good followup question for the teacher to ask? **Very bad**	Very bad question. The answer to the question is obvious even without knowing the complete text.
What was the woman doing?	Is this a good followup question for the teacher to ask? **Very bad**	Very bad question. This question is too general and wouldn't help the student provide more information.
What was the woman wearing?	Is this a good followup question for the teacher to ask? **Very bad**	Very bad question. This question asks about information which isn't present in the complete text.

Figure 7: Task examples (that are originally attached to the guidelines) for the task of rating follow-up questions.



Task

Complete text: A man holding a pink cooler is talking to an older woman.

Teacher: What do you remember about the text?

Student: A man talks to a woman.

Teacher: <Followup question with the goal of getting the student to provide more information>

Followup question	Rating	Notes(Optional)
A man with a pink cooler is talking to what kind of woman?	Is this a good followup question for the teacher to ask? N/A Very bad Bad OK Good Very good	Optional
What age is the woman?	Is this a good followup question for the teacher to ask? N/A Very bad Bad OK Good Very good	Optional
What color is the cooler?	Is this a good followup question for the teacher to ask? N/A Very bad Bad OK Good Very good	Optional
What is the man doing with his hands?	Is this a good followup question for the teacher to ask? N/A Very bad Bad OK Good Very good	Optional

General comments (optional)

comments

Figure 8: The user interface of the human annotator task of rating follow-up question.

Source text **Student description**	**A child plays with her father's boots.** **A child is playing.**
Step 1	What does she do with them?
Step 2	What does the child do with her father's boots?
Step 3	What does the child play with?
Human	What is the child playing with?
Source text **Student description**	**Two men work outside polishing shoes.** **Some men are polishing shoes.**
Step 1	What are the two men doing to the shoes?
Step 2	Who works outside to polish shoes?
Step 3	Where do the men work?
Human	How many men are there?
Source text **Student description**	**A boy dressed in a plaid kilt with a brown hat wields a long pole.** **A boy has and object in his hands.**
Step 1	Aside from the kilt, what brown item does the boy wearing it wear?
Step 2	What color is the hat the boy is wearing?
Step 3	What type of garment is the boy wearing?
Human	What does the boy wear on his body?
Source text	**A man in a white shirt and baseball hat is pushing a cart carrying several bags on a street.**
Student description	**A man is walking outside.**
Step 1	What is the man pushing a cart wearing?
Step 2	Where is the man pushing a cart with bags?
Step 3	What is the man in the picture wearing?
Human	What is the man wearing?

Table 3: Example GFQs from our different models: Step 1, Step 2, Step 3 and Human.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section (after Section 6: Conclusions)

☑ A2. Did you discuss any potential risks of your work?
Ethics and Impact section (after Section 6: Conclusions)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5.

☑ B1. Did you cite the creators of artifacts you used?
Provided a citation to the SNLI dataset and SQUAD.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix E.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Appendix E.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix G.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Appendix E.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.

C ☑ Did you run computational experiments?

Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C & D

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix C

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 5.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix B.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix B.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B.