Chapter 1: See Other File

Chapter 2: See Other File

Chapter 3: See Other File

Chapter 4: See Other File

Chapter 5: Crowd-Sourcing Non-Experts for Data

Chapters **??** and **??** use **automation** to solve a task; however, some data cannot be automatically **generated** from templates and require human assistance. One cost-efficient, scalable pool for human input are **crowd-sourcing** platforms (Background Section **??**), specifically Mechanical Turk (Buhrmester et al., 2011). We summarize a data collection project, CANARD, that uses non-**expert** workers to rewrite trivia questions.

Conversational question answering (CQA) questions differ from machine reading comprehension (MRC) ones in format (Background Section **??**); however, CQA questions can be rewritten as stand-alone MRC questions to **generate** additional training data. We reduce challenging, interconnected CQA examples to independent, stand-alone MRC to create CANARD—**C**ontext **A**bstraction: **N**ecessary **A**dditional **R**ewritten **D**iscourse—a new dataset[1] that rewrites QUAC (Choi et al., 2018) questions. We **crowd-source** context-independent paraphrases of QUAC questions and use the paraphrases to train and evaluate question-in-context rewriting. In the process, we observe the behavior of crowd users and the quality of their output.

Section 5.1 constructs CANARD, a new dataset of question-in-context with corresponding context-independent paraphrases. Section 5.2 analyzes our rewrites

[1] `http://canard.qanta.org`

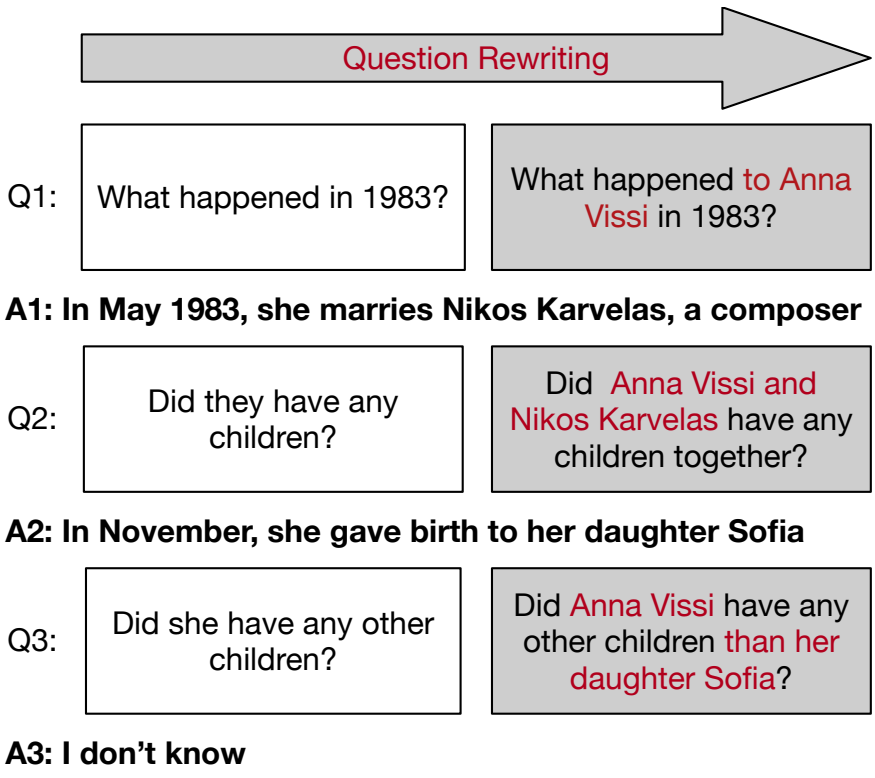

Figure 5.1: Question-in-context rewriting task. The input to each step is a question to rewrite given the dialog history which consists of the dialog utterances (questions and answers) produced before the given question is asked. The output is an equivalent, context-independent paraphrase of the input question.

(and the underlying methodology) to understand the linguistic phenomena that make CQA and using crowd-sourcing for **generation** difficult.

5.1 Dataset Construction

We elicit paraphrases from human crowdworkers to make previously context-dependent questions *unambiguously* answerable. Through this process, we resolve difficult coreference linkages and create a pair-wise mapping between ambiguous and

Characteristic	Ratio
Answer Not Referenced	0.98
Question Meaning Unchanged	0.95
Correct Coreferences	1.0
Grammatical English	1.0
Understandable w/o Context	0.90

Table 5.1: Manual inspection of 50 rewritten context-independent questions from CANARD suggests that the new questions have enough context to be independently understandable.

context-enriched questions. We derive CANARD from QUAC (Choi et al., 2018), a sequential question answering dataset about specific Wikipedia sections. QUAC uses a pair of workers—a "student" and a "teacher"—to ask and respond to questions. The "student" asks questions about a topic based on only the title of the Wikpedia article and the title of the target section. The "teacher" has access to the full Wikipedia section and provides answers by selecting text that answers the question. With this methodology, QUAC gathers 98k questions across 13,594 conversations. We take their entire dev set and a sample of their train set and create a custom JavaScript task in Mechanical Turk that allows workers to rewrite these questions. JavaScript hints help train the users and provide automated, real-time feedback.

We provide workers with a comprehensive set of instructions and task examples. We ask them to rewrite the questions in natural sounding English while

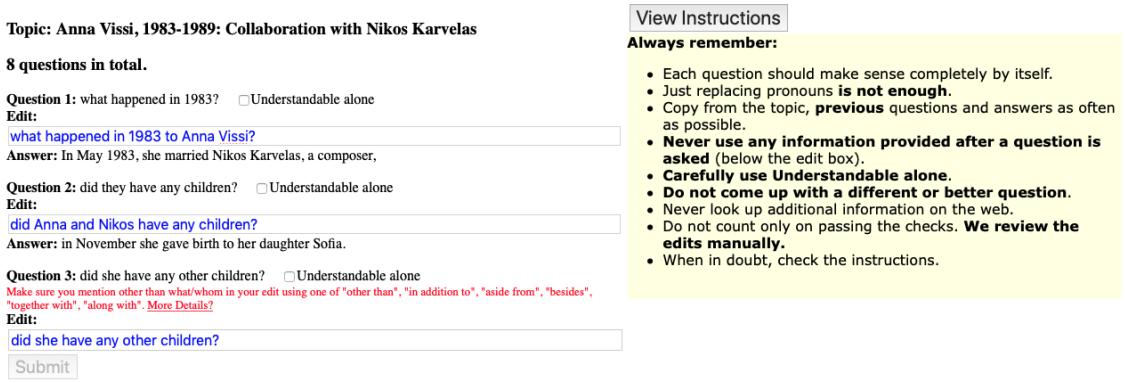

Figure 5.2: The interface for our task guides workers in real-time.

preserving the sentence structure of the original question. We discourage workers from introducing new words that are unmentioned in the previous utterances and ask them to copy phrases when appropriate from the original question. These instructions ensure that the rewrites only resolve conversation-dependent ambiguities. Thus, we encourage workers to create minimal edits.

We display the questions in the conversation one at a time, since the rewrites should include only the previous utterance. After a rewrite to the question is submitted, the answer to the question is displayed. The next question is then displayed. This repeats until the end of the conversation. Figure 5.2 displays the full set of instructions and the data collection interface.

We apply quality control throughout our collection process, given the known **generation** issues (Background Section **??**). During the task, JavaScript checks automatically monitor and warn about common errors: submissions that are abnormally short (e.g., 'why'), rewrites that still have pronouns (e.g., 'he wrote this album'), or ambiguous words (e.g., 'this article', 'that'). Many QUAC questions ask

about 'what/who else' or ask for 'other' or 'another' entity. For that class of questions, we ask workers to use a phrase such as 'other than', 'in addition to', 'aside from', 'besides', 'together with' or 'along with' with the appropriate context in their rewrite.

We gather and review our data in batches to screen potentially compromised data or low quality workers. A post-processing script flags suspicious rewrites and workers who take and abnormally long or short time. We flag about 15% of our data. *Every* flagged question is manually reviewed by one of the authors and an entire HIT is discarded if one is deemed inadequate. We reject 19.9% of submissions and the rest comprise CANARD. Additionally, we filter out under-performing workers based on these rejections from subsequent batches. To minimize risk, we limit the initial pool of workers to those that have completed 500 HITs with over 90% accuracy and offer competitive payment of $0.50 per HIT.

We verify the efficacy of our quality control through manual review. A random sample of fifty questions sampled from the final dataset is reviewed for desirable characteristics by a native English speaker in Table 5.1. Each of the positive traits occurs in 90% or more of the questions. Based on our sample, our edits retain grammaticality, leave the question meaning unchanged, and use pronouns unambiguously. There are rare occasions where workers use a part of the answer to the question being rewritten or where some of the context is left ambiguous. These infrequent mistakes should not affect our models. We provide examples of failures in Tables 5.2 and 5.4.

We use the rewrites of QUAC's development set as our test set (5,571 question-

ORIGINAL: Was this an honest mistake by the media?
REWRITE: Was the claim of media regarding Leblanc's room come to true?
ORIGINAL: What was a single from their album?
REWRITE: What was a single from horslips' *album*?
ORIGINAL: Did they marry?
REWRITE: Did Hannah Arendt and Heidegger marry?

Table 5.2: Not all rewrites correctly encode the context required to answer a question. We take two failures to provide examples of the two common issues: Changed Meaning (top) and Needs Context (middle). We provide an example with no issues (bottom) for comparison.

in-context and corresponding rewrite pairs) and use a 10% sample of QUAC's training set rewrites as our development set (3,418); the rest are training data (31,538).

5.2 Dataset Analysis

We analyze our discuss our datasets with automatic metrics. We compare our dataset to the original QUAC questions and to automatically **generated** questions by a simple seq2seq model (Background Section **??**). Then, we manually inspect the sources of rewriting errors by the model. Further improvements for the ASR dataset and CANARD are possible.

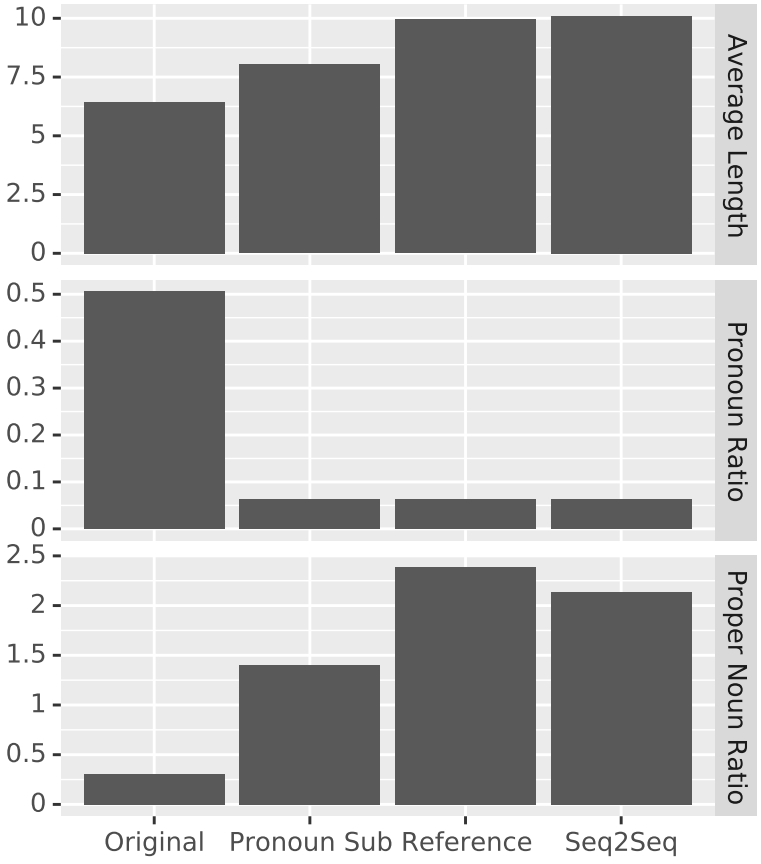

Figure 5.3: Human rewrites are longer, have fewer pronouns, and have more proper nouns than the original QUAC questions. Rewrites are longer and contain more proper nouns than our Pronoun Sub baseline and trained Seq2Seq model.

5.2.1 Anaphora Resolution and Coreference

Our rewrites are longer, contain more nouns and less pronouns, and have more word types than the original data. Machine output lies in between the two human-generated corpora, but quality is difficult to assess. Figure 5.3 shows these statistics. We motivate our rewrites by exploring linguistic properties of our data. Anaphora resolution and coreference are two core NLP tasks applicable to this dataset.

Label	Text
QUESTION	How long did he stay there?
REWRITE	How long did Cito Gaston stay at the Jays?
	Cito Gaston
	Q: What did Gaston do after the world series? ...
HISTORY	**Q:** Where did he go in 2001?
	A: In 2002, he was hired by the Jays as special assistant to president and chief executive officer Paul Godfrey.

Table 5.3: An example that had over ten flagged proper nouns in the history. Rewriting requires resolving challenging coreferences.

Pronouns occur in 53.9% of QUAC questions. Questions with pronouns are more likely to be ambiguous than those without any. Only 0.9% of these have pronouns that span more than one category (e.g., 'she' and 'his'). Hence, pronouns within a single sentence are likely unambiguous. However, 75.0% of the aggregate history has pronouns and the percentage of mixed category pronouns increase to 27.8% of our data. Therefore, pronoun disambiguation potentially becomes a problem for a quarter of the original data. An example is provided in Table 5.3.

Approximately one-third of the questions generated by our pronoun-replacement baseline are within 85% string similarity to our rewritten questions. That leaves two-thirds of our data that cannot be solved with pronoun resolution alone.

5.3 Conclusion

A limitation of generalist **crowd-sourcing** is the inability to automatically quality control **generated** data. Our work requires *manual* analysis of each sentence submitted by the crowd; this is time-intensive and subject to error. Additionally, it requires real-time task monitoring and user exclusion as otherwise malicious users can quickly contribute a large part of your crowd-sourced task. There is no full-proof way to ensure quality in tasks involving crowd-sourcing **generation**. However, this method generates more diverse and lengthy sentences than comparable **automation** projects (Chapters **??** and **??**) One way to handle the quality control issue is by using an **expert** for both **generation** and for quality assessment (Chapter **??**).

Original	**Rewrite**	**Issue**
Was the Belmont a close race as well?	Was the Belmont a close race as well?	Identical
Did they argue?	DID JOHNSON AND BIRD ARGUE?	Capitalization
Did it affect future games?	Were future game affected?	Grammar Change
did he have other musical films?	did Frank Sinatra have other musical films?	Incomplete Expansion
Any others?	Besides Menswear Designer of the Year any others?	Nonsensical Expansion
How many copies of the album was sold?	Besides shipping 1,000,000 units how many copies of the album where sold?	Nonsensical Expansion

Table 5.4: UPDATE TO BE SIMILAR TO OTHER TABLE OR REMOVE. Additional examples of **generated** rewrites by crowd-workers. Manual inspection is necessary to ensure these are not deemed accurate question rewrites.

Bibliography

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on psychological science: a journal of the Association for Psychological Science*, 6 1:3–5.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1096–1104.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *Proceedings of the Association for Computational Linguistics*.

Frank Wessel and Hermann Ney. 2004. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*.