A Crowdsourceable Protocol for Annotating Multi-Hop QA with Reasoning Steps

Naoya Inoue[1,2] Pontus Stenetorp[3,2] Kentaro Inui[1,2]
[1]Tohoku University [2]RIKEN Advanced Intelligence Project [3]University College London
{naoya-i,inui}@ecei.tohoku.ac.jp pontus@stenetorp.se

1 Introduction

Discourse processing (DP) is a long-standing goal in artificial intelligence. DP is beneficial for a wide variety of applications, such as question qnswering [21], document classification [13], and essay scoring [19]. This paper explores a new direction for discourse processing—DP models *that perform logically correct reasoning*.

This direction of research is important for three reasons. First, encouraging DP models to perform correct reasoning is expected to work as a *regularizer*. Recent studies [14] show that the current reading comprehension (RC) models, an instance of DP models, are vulnerable to adversarial inputs. We speculate that the current RC models overfit to a given benchmark, because they are tuned only towards predicting correct answers, neglecting their reasoning they make. RC models that performs correct reasoning will be more robust to unseen, noisy examples outside of training/test examples.

Second, DP models will require less training data. Without explicit supervision of reasoning, machine learning models try to induce complex prediction rules for DP by generalizing training instances. This is particularly challenging in the limited availability of training data; guiding reasoning as explicit supervision will be helpful [22]. Recent studies show that exploiting human rationales in sentiment analysis improves the prediction performance in low resource settings [2, 9].

Finally, DP models will be more transparent. Machines' ability to explain its own prediction is proven to be important in applications. Zhang et al. [6] show that the user engagement of recommendation systems increases when they are provided with explanations as to why the recommendation is given. The importance of interpretability of models is also extensively discussed in the Machine Learning community [7].

There are three approaches related to DP models and reasoning. First, there is a large body of work on analyzing the behaviour of RC models, motivated by the question that RC models understand natural language in the real sense [? 17, etc.]. Second, there are NLP tasks that require models to output its own prediction, including QA tasks [22], fact-checking tasks [18], and interpretable textual similarity tasks [1]. Third, several datasets are annotated with explanations (i.e. reasoning), ranging from science QA dataset [10, 12, 11], entailments [5] to argumentative texts [4, 3, 8]. Nevertheless, none of them create a high-quality, large-scale corpus annotated with reasoning, which drives the research of DP models enhanced with reasoning.

Given the background, we create a new corpus annotated with reasoning. This poses the following challenges. First, reasoning can be arbitrary. How can we design representations of reasoning that can be consistently annotated and used as supervision to correctly guide machines' predictions? (C1) Second, annotating reasoning is costly. However, the corpus needs to be large-scale and of high quality for training and evaluating DP models. How can we ensure the scalability while maintaining the quality? (C2)

This paper provides a solution to these challenges. For C1, we limit our problem scope to multi-hop QAs that require machines to combine several relational facts between entities to derive an answer. We then define reasoning as a sequence of relational facts between entities represented by natural language sentences. For example, given the statement "*Machu Picchu is in Peru.*", we represent reasoning as ("*Machu Picchu is in the Andes Mountains.*", "*The Andes Mountains are in Peru.*"). For C2, we develop a crowdsourceable annotation protocol for reasoning and carefully design a step-by-step annotation interface. The contributions of this paper can be summarized as follows:

- We show how to cast the complicated, reasoning annotation task into an annotation task that can be done by non-expert workers.
- The proposed reasoning representation is richer than previously proposed representations (i.e. sentence labels [22], or single reasoning steps [9]).
- We apply the developed crowdsourcing task to 2,000 QA pairs on WikiHop [20][1], a popular multi-hop QA dataset. We make the corpus of reasoning annotations publicly available.[2]

2 Crowdsourced reasoning annotation

2.1 Key idea

We consider multi-hop QA [20, 22] as a testbed for reasoning annotation. In multi-hop QA, unlike regular QA, machines need to perform reasoning by combining several relational facts between entities to derive an answer

[1] https://qangaroo.cs.ucl.ac.uk/
[2] https://github.com/cl-tohoku/reasoning_annotation

(i.e. multi-hop). Suppose the question "*What is the nationality of Trump?*" and the following passages P_1, P_2:
• p_1: Trump was born and raised in New York City.
• p_2: New York City is a city in the US.
To answer the question, combining the facts about Trump and New York City is needed.

Using multi-hop QA as a benchmark has several advantages. First, it offers a richer set of problems that require machines to perform reasoning. Second, reasoning representation is suitable as a first exploration of DP with reasoning. As we see later, reasoning in multi-hop QA can be represented as a sequence of relations between entities. This allows (i) machines to easily evaluate their reasoning and to be guided, and (ii) non-expert workers to easily annotate reasoning.

2.2 Annotation scheme

Given a relational statement of entities $r(e_i, e_j)$ and n related passages $p_1, p_2, ..., p_n$, we annotate them with reasoning as to why the statement is true *solely according to the passages*. We define reasoning as an n minimal chain of relational statement to derive an answer, i.e. $(r_1(e_i, e_1), r_2(e_1, e_2), ..., r_{n-1}(e_{n-1}, e_n), r_n(e_n, e_j))$. We call each statement a *reasoning step*.

Suppose the relational statement "*Trump's nationality is the US*" along with the above passages p_1, p_2. We annotate them with the following two reasoning steps:
• Step 1: ***Trump** was born in **New York City**.*
• Step 2: ***New York City** is located in **the US**.*
On the other hand, suppose p_1 were "Trump owned rental housing in New York City." We do not annotate them with reasoning, because we cannot conclude the truth of the statement, solely based on these passages. We simply mark them as UNREACHABLE.

For relational representations r_i, we employ natural language to investigate type of reasoning steps required in multi-hop QA. Previous works on annotating explanations [18, 22, etc.] formulate the explanation annotation task as a sentence selection task in a given passage. Our annotation provides richer information than these formulations, because original sentences may not contain "minimal" information relevant to reasoning.

2.3 Crowdsourcing interface

To scale up the annotation, we implement the annotation scheme via crowdsourcing (CS). Recent studies show that CS is a powerful tool for creating a large-scale dataset for NLP [22, 5, etc.]. One big obstacle of using CS is in the difficulty of controlling the quality of annotation results. CS requiring crowdworkers to input a free-form text is particularly hard, because it is not easy to validate their input given a free text. We thus carefully design a step-by-step annotation interface.

According to the annotation scheme, we first give crowdworkers an instruction with examples. In a welcome message, we try to motivate workers by saying that they are involved in educating artificial intelligence. We also make sure that they judge the truth of statements *solely based on given passages, not based on their own knowledge.*

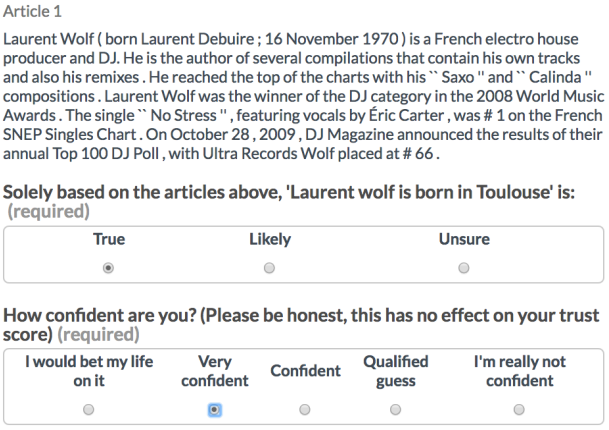

Figure 1: Crowdsourcing interface: selection task.

The annotation interface follows the instruction. Overall, our CS task asks crowdworkers two types of jobs. First, we describe the *selection task* illustrated in Figure 1. Given a relational statement (i.e. "*Laurent Wolf is born in Toulouse*") and the first passage (i.e. Article 1), workers are required to judge whether the statement can be derived from the passage at three grades. To filter out unreliable annotation results, we also ask how confident they are. If a worker selects Unsure, we show the statement again with *two passages* and ask the same question.

If a worker selects True or Likely, we proceed to the *writing task*. We ask workers to write reasoning steps they made to derive the statement, based on given passages. We encourage workers to summarize their reasoning steps, not just to extract sentences from the passages. For the first and last textbox, we use a textform validator that requires the textbox to contain e_1 (i.e. *Laurent Wolf*) and e_2 (i.e. *Toulouse*).

In our preliminary experiment, we found that this "step-by-step" task is important. We speculate that workers are overwhelmed when all passages are given at a time (up to three passages in this study).

2.4 Settings

To ensure the quality of annotation results, we tuned several CS parameters. Because the reasoning step annotation is a time-consuming task, we pay 10 cents as a reward to crowdworkers per one instance. In our preliminary experiments, we found that some crowdworkers are lazy, e.g., entering random sentences such as "asdjlkkjsdfkj" or doing copy & paste from some sentences in the instruction. We thus set the minimum work time to 30 seconds, where workers who completed their work less than 30 seconds are judged as untrusted, and kicked out from the task. The workers are shown two instances per page, and allowed to submit their work after they finish two instances.

2.5 Heuristic postprocessing

Due to the nature of CS, the results are expected to be noisy (e.g. repeating the same sentences in textboxes). We thus heuristically discard annotations if one of the following conditions is satisfied: (i) one of reasoning

Hop	# resp.	Example of worker responses for reasoning steps
1	2,159 / 1,138	Statement: Mellerstain house is located in Scottish borders [1] mellerstain house is 13 kilometers north of kelso in the scottish borders
2	1,052 / 486	Statement: Krisztina pigniczki is a citizen of Hungary [1] Krisztina Pigniczki was born in Makó [2] Makó is a town in Csongrád County, in southeastern Hungary
3	458 / 162	Statement: A genre of Snacka om nyheter is Panel game [1] Snacka om nyheter was the Swedish version of the BBC series Have I Got News for You [2] Have I Got News for You is loosely based on the BBC Radio 4 show "The News Quiz" [3] The News Quiz is a British topical panel game
Unr.	6,331 / 4,538	
Overall	10,000 / 6,324	

Table 1: Distribution of worker's responses and example of annotated reasoning steps (before the slash: raw annotations, after the slash: annotations with the postprocessing.

steps is exactly the same as a statement; (ii) for # hops ≥ 2, at least one pair of reasoning steps have word overlap ratio ≥ 0.9, (iii) at least one confidence value is "Unsure", or (iv) one of reasoning step exactly matches with a sentence in an article.

3 Evaluation

3.1 Settings

We implemented the proposed protocol on FigureEight (a.k.a CrowdFlower)[3], a widely used CS platform. To see how reasoning varies across workers, we hire 5 crowdworkers per one instance.

3.2 Dataset

There are a few choices of datasets for multi-hop QA [20, 22, etc.]. This study uses WikiHop [20], as it has been widely used as a benchmark for multi-hop QA.[4] We randomly sampled 2,000 instances from 5,129 instances in the development set. For the CS task, we manually converted QA instances into natural language statements.

For each question, WikiHop provides a set of supporting documents. Based on the distant supervision assumption [15], WikiHop collects a set of Wikipedia articles that bridges an entity in a question (i.e. e_i) and an entity e_j in an answer, where the link between articles is given by a hyperlink. The number of supporting documents ranges from 2 to 3. In the development set, there are 892 instances with 2 supporting documents, and 1,108 instances with 3 supporting documents.

3.3 Results and discussion

3.3.1 Task evaluation

It took about 4 hours to finish the whole task. It cost $1,208 (including transaction fees).

The survey of our CS task by 46 participants indicates that (i) our instructions are considered almost clear by crowdworkers (ratings: 3.6/5) and (ii) the pay (10 cents per instance) is considered reasonable by workers (ratings: 3.9/5). It also indicates that the task was considered difficult (ratings: 3.0/5) partly because of the length of the articles. We are planning to improve the interface by highlighting entities shared by adjacent articles.

[3] https://www.figure-eight.com/

[4] The leaderbaord https://qangaroo.cs.ucl.ac.uk/leaderboard.html received 17 submissions, as of January 16th, 2019.

Hop	# hops		Reasoning	
	Strict	Lenient	Strict	Lenient
1	.00 / .00	.90 / **1.00**	.00 / .00	.50 / **1.00**
2	.30 / **.80**	.60 / **.90**	.20 / **.70**	.70 / **.80**
3	.30 / .30	.60 / **.90**	.30 / **.40**	.50 / **.90**
Unr.	**.80** / .70	**.80** / .60	-	-
Overall	.35 / **.45**	.73 / **.85**	.35 / **.37**	.57 / **.90**

Table 2: Precision of raw annotations (before the slash) and annotations with heuristic postprocessing (after the slash).

3.3.2 Responses

Table 1 shows the distribution of hop and reasoning steps annotated by crowdworkers. After the heuristic postprocessing, the total number of annotations is 6,324, where 1,982 instances have at least one annotation.

To estimate the quality of crowdsourced annotation, we randomly sampled 10 annotations from each hop and manually checked whether (i) the number of hops and (ii) annotated reasoning steps are valid or not. Table 2 shows the results of the manual analysis. In the *strict evaluation*, we judged if the judgement of a crowdworker is *strictly* based on given passages. In the *lenient evaluation*, we allow crowdworkers to use a bit of implicitly stated common knowledge. For example, if the passage mentions that "`X is a Moroccan football player`", then we allow them to conclude that "X's nationality is Morocco", even though it is not mentioned that "if someone is a moroccan football player, their nationality is morocco". We believe that these reasoning steps are still useful for guiding machines' prediction.

The results indicate that the heuristic postprocessing helps to mine high-quality annotation results, reducing the annotation results from 10,000 instances to 6,324 instances. We employ the postprocessed corpus in the remaining analysis.

3.3.3 Agreement study

To get further insights on the quality of CS annotation, we conducted an agreement study.[5] For the number of hops, we simply calculated the percentage of agreed number of hops among annotations. For the reasoning steps, we calculated BLEU-4 [16] in a pairwise manner and averaged them.[6]

[5] We consider only 1,861 instances that have more than one annotations after the postprocessing.

[6] Only pairs of instances with the same reasoning steps are considered.

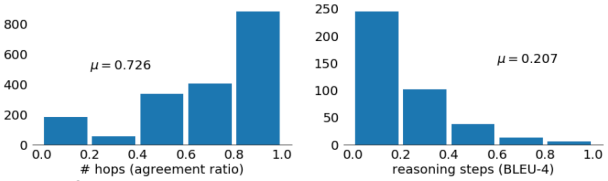

Figure 2: Histogram of agreement ratio of # hops and BLEU-based agreement measure of reasoning steps.

Figure 2 shows the histogram of the agreement values. For the number of hops, the results indicate fairy high agreement. To see how workers disagree, we manually analyzed 20 instances with worst agreement ratio. We found that the majority of disagreements (15/20) come from "unreachable v.s. reachable". However, our manual inspection found that crowdworkers who say "reachable" are reliable—12/15 annotations can be actually judged as reachable (in the lenient criterion). Surprisingly, despite the text input task, annotated reasoning steps are also of good quality. We found 10/15 can be judged as valid (in the lenient criterion). Regarding the disagreement among reachable annotations (5/20), we also found that all of them (both # hops and reasoning steps) can be judged as valid (in the lenient criterion).

Overall, the results indicate that workers who wrote reasoning steps are quite reliable. We speculate that such workers are motivated, because they could also earn money even without writing reasoning steps.

On the other hand, the reasoning agreement indicates that reasoning steps vary across workers. To analyze the cause of disagreement of reasoning steps, we have looked at 10 instances with the worst BLEU values. We found that some workers inserted some descriptive phrases such as "*the article said that...*" "*the article does not mention...*" into a text box. In other cases, one worker simply extracted a sentence from a Wikipedia article, and the other summarized their reasoning steps. We believe that these disagreements do not have a significant impact on the quality of the corpus.

3.3.4 Final corpus

A common strategy for aggregating CS annotations relies on agreement (e.g. majority voting). However, the analysis in Sec. 3.3.3 found that the central factor is not agreement, but whether crowdworkers provide with reasoning steps or not. We thus take the maximum number of hops provided by crowdworkers for aggregation.

The final distribution of the number of hops is: (i) # hops = 1: 0.29 (574/1,982), (ii) # hops = 2: 0.19 (377/1,982), (iii) # hops = 3: 0.07 (145/1,982), and (iv) unreachable: 0.45 (886/1,982).

4 Conclusions

We have explored a method to create a large-scale corpus of reasoning. Taking multi-hop QA as a testbed, we carefully design the crowdsourcing interface. Our experiments demonstrate that the proposed CS protocol produces a large-scale, high-quality corpus of reasoning on top of WikiHop [20]. We make the corpus of reasoning annotations publicly available.

One immediate future work is to expand the annotation to 43,738 WikiHop training instances, or other multi-hop QA datasets such as HotpotQA [22]. This enables us to conduct a large-scale study of discourse processing models enhanced with reasoning (e.g. investigating robustness to adversarial examples).

Acknowledgements

We thank Johannes Welbl at University College London for their assistance with obtaining supporting documents in WikiHop. This work was partially supported by JST-Mirai Program JPMJMI17C7, and JST CREST Grant Number JPMJCR1513, including AIP challenge.

References

[1] Eneko Agirre, Aitor Gonzalez-agirre, Inigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. SemEval-2016 Task 2 : Interpretable Semantic Textual Similarity. *SemEval*, pages 512–524, 2016.

[2] Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving Machine Attention from Human Rationales. pages 1903–1913, 2018.

[3] Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. Enriching argumentative texts with implicit knowledge. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10260 LNCS:84–96, 2017.

[4] Filip Boltuzic and Jan Šnajder. Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates. *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, 2016.

[5] Oana-maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI : Natural Language Inference with Natural Language Explanations. *32nd Conference on Neural Information Processing Systems (NIPS)*, (NeurIPS):1–13, 2018.

[6] Yongfeng Zhang Chen and Xu. Explainable Recommendation: A Survey and New Perspectives. Technical report, 2018.

[7] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. (Ml):1–13, 2017.

[8] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. pages 1930–1940, 2017.

[9] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. Training Classifiers with Natural Language Explanations. pages 1884–1895, 2018.

[10] Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exams. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, 2016.

[11] Peter A Jansen. Multi-hop Inference for Sentence-level TextGraphs: How Challenging is Meaningfully Combining Information for Science Question Answering? pages 12–17, 2018.

[12] Peter A. Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison. WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-Hop Inference. 2018.

[13] Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. Dynamic Entity Representations in Neural Language Models. pages 1831–1840, 2017.

[14] Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. 2017.

[15] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. *Procs. of ACL-IJCNLP*, 2(August):1003, 2009.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Procs. of ACL*, pages 311–318, Philadelphia, Pennsylvania, USA, 7 2002. Association for Computational Linguistics.

[17] Saku Sugawara and Akiko Aizawa. What Makes Reading Comprehension Questions Easier? 2018.

[18] James Thorne and Andreas Vlachos. Automated Fact Checking: Task formulations, methods and future directions. pages 3346–3359, 2018.

[19] Henning Wachsmuth and Benno Stein. A Universal Model for Discourse-Level Argumentation Analysis. *ACM Transactions on Internet Technology*, 17(3):1–24, 2017.

[20] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. 2017.

[21] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. 2015.

[22] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. 2018.