Appendix

For the crowdsourcing task described in Sec. 4.2.1, we have a gold set annotated by experts, based on which CrowdFlower has two types of quality controls: First, each crowdsourcer has to pass a qualifying test, which is randomly picked from the gold set, with accuracy higher than 70% before the crowdsourcer is allowed on a job; one is also tested on the gold set randomly throughout the process without notice and once the accuracy drops below 70%, the crowdsourcer is kicked out automatically and his or her annotations are all considered tainted and not used. Finally kept are those annotations from crowdsourcers who survive in the end (so called Trusted Contributors). **In our case, trusted contributors had 87% accuracy on our gold set**, as reported by CrowdFlower.

Accuracy on the gold set reflects the crowdsourcers' level of understanding of the job owner's intent. Another quality metric is the level of agreement among themselves. Since more than two annotators are involved in crowdsourcing, Cohen's Kappa cannot be applied to crowdsourced data. Instead, CrowdFlower adopts a default IAA metric called WAWA (Worker Agreement with Aggregate). WAWA indicates the average number of crowdsourcers' responses agreed with the aggregate answer for each question. For example, if N individual responses were obtained in total, and n of them were correct when compared to the aggregate answer, then WAWA is simply n/N. **The WAWA score of the proposed new dataset was 78%.**

In terms of the cost of the crowdsourcing job, we paid \$0.01 for every individual response and the total cost was approximately \$500 (including overhead fees). The actual annotation time was about 15 hours for the entire dataset to be finished.