Does History Help?
An Experiment on How Context Affects Crowdsourcing Dialogue Annotation

Elnaz Nouri
Institute for Creative Technologies, Computer Science Department, University of Southern California
enouri@usc.edu

Abstract

Crowds of people can potentially solve some problems faster than individuals. Crowd sourced data can be leveraged to benefit the crowd by providing information or solutions faster than traditional means. Many tasks needed for developing dialogue systems such as annotation can benefit from crowdsourcing as well. We investigate how to outsource dialogue data annotation through Amazon Mechanical Turk. We are in particular interested in empirically analyzing how much context from previous parts of the dialogue (e.g. previous dialogue turns) is needed to be provided before the target part (dialogue turn) is presented to the annotator. The answer to this question is essentially important for leveraging crowd sourced data for appropriate and efficient response and coordination. We study the effect of presenting different numbers of previous data (turns) to the Turkers in annotating sentiments of dyadic negotiation dialogs on the inter annotator reliability and comparison to the gold standard.

Problem and Contribution

Annotating dialogues is useful for building dialogue systems based on statistical models and verifying theories of dialogue. Since the mid 1990s, increasing effort has gone into putting semantics and discourse research on the same empirical footing as other areas of computational linguistics (Artstein et al., 2008). Annotating dialogues is critical for training supervised classifiers but doing it manually in the lab is time consuming and expensive. Crowdsourcing dialogue annotation tasks to a large population of online workers seems to be a good solution for bringing down the costs and time needed for accomplishing these tasks. Crowdsourcing annotation tasks has been shown to help with these aspects. Several recent papers have studied the use of annotations obtained from Amazon Mechanical Turk (Su et al., 2007; Kaisser et al., 2008; Kittur et al., 2008; Sheng et al., 2008; Snow et al., 2008; Sorokin et al. , 2008). (Hsueh et al., 2009) compares the quality of the annotated data from experts in the research lab and non-experts recruited on the web. Consistent with the results reported in (Sheng et al., 2008) they show that using multiple noisy annotations from different non-experts can still be very useful for modeling.

Dialogue data is sequential by nature and that differentiates the process for crowdsourcing dialogue tasks from other categorization or annotation tasks (Parent et al.,2011, Suendermann et al., 2013). Considering the low payment range to online mechanical workers it seems unreasonable to give them the whole conversation that has happened prior to the item that is to be annotated. Giving a partial history of what has occurred before the target item seems to be a reasonable solution. However it's not known how big a window would be enough for this purpose and how the length of the history of the conversation prior to the target item would affect the reliability of the annotation task. Our goal is to study the amount of prior context needed for crowdsourcing dialogue annotation tasks.

As an empirical test-bed for addressing the above question we consider the problem of annotating the sentiments on a negotiation dataset. (Delany et al., 2011) proposes that crowdsourcing for labeling emotional assets but this work along other current work available on annotation of dialogues (Parent et al. 2011, Suendermann et al. 2013) do not address the questions that we are trying to answer.

Experiment Details

We performed a sentiment annotation task on transcription of negotiation dialogues. Three negotiation

dialogues were analyzed for this report due to time and resource constraints. This work is a preliminary analysis of this dataset. The details of our current methodology and analysis are provided here.

Dataset

The "Farmers Market" negotiation data was collected in USC's Marshall School of business, by Peter Carnevale, based on a negotiation task over multiple objects. This data set was also used in (Nouri, 2013). The dataset consists of 41 dyadic sessions.

Sentiment Annotation Task

The Mechanical Turk annotators were asked to judge the sentiment expressed by each turn in the dialogue based on the criteria presented in table 1:

Emotion Tag	Score	Emotion Embodied
Strongly positive	2	extremely happy or excited toward the topic
Positive	1	generally happy or satisfied, but the emotion wasn't extreme.
Neutral	0	Not positive or negative
Negative	-1	perceived to be angry or upsetting toward the topic, but not to the extreme
Strongly Negative	-2	extremely negative toward the topic

Table 1.Sentiment Annotation Schema

Crowdsourcing

The annotators were recruited on Amazon Turk. They were paid $0.02 for judging the sentiment on each turn of the dialogue. For each annotation instance 5 workers judged the sentiment of the dialogue turn.

Sequential History Window

Each annotation task consisted of judging the sentiment of one single turn in the conversation. However the experiment was run in 7 rounds. The only difference between the rounds was the number of preceding turns being shown before the target dialogue turn that was being judged. The annotators were provided with 0, 1, 2, 3, 4, 5 and 6 (size of the dialogue history window) previous turns in the dialogue before being asked to annotate the turn that followed the provided sequence.

Results

We present two types of analysis in order to measure the differences in the quality of collected annotations. Inspired by the method used in (Hsueh et al. 2009), the first method compares the distance of average annotation scores by the Turk annotators with the Gold baseline annotation score of the turns. The Gold baseline scores was calculated by having one expert annotator annotate the sentiment when presented with the complete dialogues. We suggest that this measure can capture how much difference the size of previous dialogue context makes on the annotation.

Table 2 shows the result of this analysis on the three sample negotiation dialogues. The numbers don't show significant differences against the number of turns presented before the target turn. We have highlighted the minimum distance for each dialogue with the sign * in the table. One might conclude that at least one turn before the target turn is needed to help the annotators. It is also possible to suggest that the number of previous turns doesn't really affect the annotation of the target turn and hence conclude that showing a big number of previous turn to the annotator might not be necessary.

Dialogue	D1	D2	D3
0 turns	0.260	0.341	0.236
1 turns	0.261	0.317*	0.228*
2 turns	0.215*	0.326	0.248
3 turns	0.299	0.349	0.313
4 turns	0.277	0.413	0.268
5 turns	0.238	0.356	0.247
6 turns	0.246	0.341	0.255

Table 2. Distance of the MTurk annotation scores (with different number of turns presented before the goal turn) from the Gold annotation for the three sample dialogs

For the second method based on (Artstein, 2008) we decided to use the Krippendorff's alpha score to measure the agreement between the five workers in annotations on the dialogues. Our hypothesis is that higher inter-annotator reliability implies more stability and can be used as an indication of the optimal number of context needed for crowdsourcing annotation tasks. The hypothesis is still to be verified by extensive empirical experiments. Table 3 shows the corresponding results based on how many turns were shown before the target turn was annotated. The tags were considered as nominal labels for measuring this score.

Number of turns before target	Krippendorff's alpha
0	0.0976
1	0.2165
2	0.1133
3	0.2431
4	0.1670
5	0.1923
6	0.1790

Table 3. Inter-annotator reliability based on Krippendorff's alpha

The highest inter annotator agreement is seen when three previous turns are shown to the Turkers before the target turn. Just as with the results of the previous method, according to this table it is possible to suggest that showing the whole dialogue or a big number of context turns to the annotators is not required and one can get reasonable results by just presenting few previous turns.

Conclusion and Future Work

Based on our current results we suspect that a history window size of 3 is likely to be the optimal number of turns needed. This is based on the assumption that higher inter-annotator reliability indicates stability. The 5 annotators have the highest agreement score having been provided with three previous turns in the dialogue. Our sample size is very small at this point and it's not possible to make strong conclusions based on the current results. More experiments are needed to make a general recommendation on this issue. The nature of the dialogues and the annotation task can also affect these results and needs to be further investigated.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555-596.

Delany, S. J., Tarasov, A., Snel, J., & Cullen, C. (2011). Using Crowdsourcing in the Rating of Emotional Speech Assets.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House. Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009, June). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing* (pp. 27-35). Association for Computational Linguistics.

Kaisser, M., Hearst, M., & Lowe, J. B. (2008). Evidence for varying search results summary lengths. In Proc. of ACL.

Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453-456). ACM.

Nouri, E., Park, S., Scherer S., Gratch, J., Carnevale, P., Morency, LP., Traum, D. (2013). *Prediction of Strategy and Outcome as Negotiation Unfolds by Using Basic Verbal and Behavioral Features*. Interspeech 2013.

Parent, G., & Eskenazi, M. (2011, August). *Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges.* In INTERSPEECH (pp. 3037-3040).

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 614-622). ACM.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). Association for Computational Linguistics.

Sorokin, A., & Forsyth, D. (2008, June). Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (pp. 1-8). IEEE.

Su, Q., Pavlov, D., Chow, J. H., & Baker, W. C. (2007, May). Internet-scale collection of human-reviewed data. In *Proceedings of the 16th international conference on World Wide Web* (pp. 231-240). ACM.

Suendermann, D., & Pieraccini, R. (2013). *Crowdsourcing for Industrial Spoken Dialog Systems. Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, 280-302.