Expert-generated vs. Crowd-sourced Annotations for Evaluating Chatting Sessions at the Turn Level

WOCHAT SHARED-TASK REPORT

Rafael E. Banchs

Institute for Infocomm Research, Singapore
rembanchs@i2r.a-star.edu.sg

Abstract. This report presents a comparative study between expert-generated and crowd-sourced annotations for evaluating human-chatbot interactions according to the guidelines in WOCHAT's shared task. Two expert annotations are compared to three crowd-sourced annotations and two different combinations of them, showing that majority voting over crowd-sourced annotations exhibits similar inter-annotator agreements to those observed between expert-generated annotations.

Keywords. Crowd-sourcing, Chatting Evaluation, Inter-annotator Agreement, Turn-level Annotations.

1 Introduction

This paper focuses on the problem of evaluating chatting sessions at the turn level by means of crowd-sourcing resources. Following the annotation guidelines defined in the WOCHAT shared task, the evaluation is performed by assigning one of three possible subjective scores (*valid* / *acceptable* / *invalid*) to each turn in a given chatting session. More details on the proposed turn-level subjective annotation scheme are provided in the shared task's Annotation Guidelines[1]

In addition to its subjective nature, this kind of annotation is an exhausting and time consuming task. Given the necessity of annotating much more data and at a much higher speed, in this second edition of the shared task a crowd-sourced annotation exercise was conducted with the objective of evaluating the quality of the resulting annotations, as well as the cost and speed of such an approach.

2 Expert Annotations and Selected Chatting Sessions

After analyzing all annotated chatting sessions from the first edition of the shared task, we found out that the largest set of sessions commonly annotated by the same two experts reduces to a subset of *16* sessions, compressing a total of *437* turns. From

[1] http://workshop.colips.org/wochat/documents/Annotation_Guidelines.pdf

all these chatting sessions, both the first and the last turns were removed as they corresponded to the standard chatbot session initiation and termination, which were always consistently evaluated as *valid* by both expert annotators. As a result, the comparative analysis presented here was conducted over a subset of *405* turns.

Two inter-annotator agreement coefficients were computed over the two sets of expert annotations: Fisher's Interclass Correlation coefficient [2] and Cohen's Kappa coefficient [1]. The resulting values were *0.5454* and *0.3736*, respectively, which are considered as *fair* agreement in both scales.

3 Turn-level Annotation HITs

In order to conduct the subjective evaluations by means of crowd-sourcing, the Amazon Mechanical Turk[2] platform was used. For collecting annotations at the turn level, the HITs (Human Intelligence Tasks) were defined in such a way that each annotator assigned one of the three possible scores to a single turn given the previous context in the chatting session. The chatting context-history length was set to three, which means that the annotator had to decide on what score to assign based on the current turn and the three previous ones. Figure 1 shows an example of HIT.

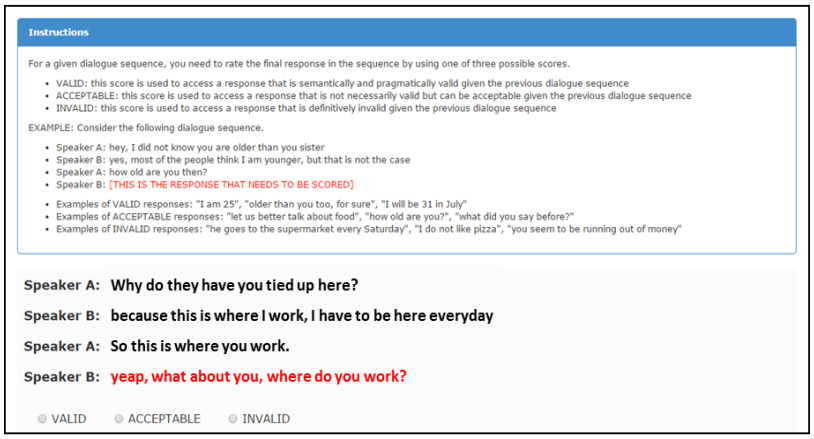

Fig. 1. Example of HIT for turn-level annotation of chatting sessions

4 Crowd-sourced Annotations

A total of *405* HITs were published in the Amazon Mechanical Turk platform for annotation. Three assignments were requested for each HIT, meaning that each turn had to be annotated by three different workers, thus collecting *1,215* annotations in total. Each annotation was paid at US$0.05 plus a US$0.01 fee charged by Amazon, for a cost of US$0.06 per annotation and a total US$72.90 for the complete exercise.

[2] https://www.mturk.com/mturk/welcome

Thirty three workers participated in the annotations, with the number of assignments per worker varying from *1* to *187*. The average number of assignments per worker was *36.82*, although the resulting distribution was highly skewed. The average time per assignment was *30* seconds, resulting on an effective hourly rate of US$6.81.

Different from expert-generated annotations, in the crowd-sourced ones all turns in a same chatting session are not consistently scored by the same annotator. According to this, it does not make sense to compute inter-annotator agreements among assignment sets. Alternatively, we looked at the distributions of inter-annotator agreement coefficients over different permutations of the three conducted assignments. To this end, a total of *30,000* permutations were randomly selected over the *405* scored turns. Kappa and Interclass Correlation Coefficient (ICC) between the permutations and the two expert annotations were computed. The resulting mean values and standard deviations are summarized in Table 1 and the distributions are depicted in Figure 2.

Table 1. Mean values and standard deviations for Kappa and ICC between expert and crowd-sourced annotations.

	Kappa Mean	**Kappa Std**	**ICC Mean**	**ICC Std**
Expert 1	*0.2918*	*0.0225*	*0.4401*	*0.0267*
Expert 2	*0.3419*	*0.0248*	*0.5121*	*0.0254*

The mean values for both, Kappa and ICC, between crowd-sourced and expert annotations are lower than the corresponding values between the two experts: ICC *0.5454* and Kappa *0.3736* (see section 2). Notice also that, in terms of inter-annotator agreement, the crowd-sourced annotations are closer to Expert 2 than Expert 1.

5 Combining Crowd-sourced Annotations

As seen in the previous section, crowd-sourced annotations are not comparable to expert-generated ones. In this section we explore whether the combination of crowd-sourced annotations can get closer to expert annotations. Two different majority voting methods are considered to combine the three scores available for each turn:

- **Method A: Majority voting with strong disagreement penalty.** If two or more scores are equal, the resulting score is set accordingly with the exception of strong disagreement cases. The strong disagreement cases (1) three different scores, (2) two *valids* and one *invalid*, (3) one *valid* and two *invalids* are all set to *acceptable*.
- **Method B: Simple Majority voting.** No strong disagreement penalization is used. The resulting score is set accordingly whenever two or more scores are equal. However, if the three scores are different, the score is still set to *acceptable*.

Table 2 presents the resulting Kappa and ICC between expert-generated and combined crowd-sourced annotations for both methods. As seen from the table, Method B performs better, producing annotations that exhibit higher inter-annotator agreement with both expert-generated ones. Figure 2 illustrates the results presented in Tables 1 and 2 for the case of ICC (Kappa, not presented here, behaves similarly).

Table 2. Kappa and ICC between expert and combined crowd-sourced annotations.

	Kappa Method A	Kappa Method B	ICC Method A	ICC Method B
Expert 1	*0.3011*	*0.3259*	*0.5236*	*0.5380*
Expert 2	*0.3979*	*0.4285*	*0.5919*	*0.5951*

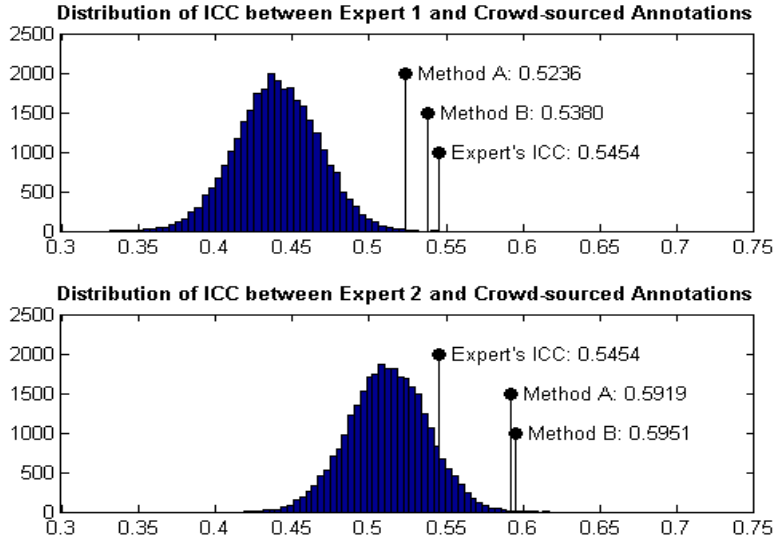

Fig. 2. Distributions of ICC between expert-generated and crowd-sourced annotations shown along with ICC values achieved by the proposed combination methods and between experts.

6 Conclusions

This report presented a comparative study between expert-generated and crowd-sourced annotations for evaluating chatting sessions at the turn-level according to the guidelines defined in the WOCHAT shared task. Two expert annotations over 16 chatting sessions were compared with three crowd-sourced annotations and two different majority voting strategies. As a result, crowd-sourced annotations exhibited lower inter-annotator agreements with expert annotations than the one between experts. However, the simple majority voting strategy over crowd-sourced annotations produced annotations that exhibit similar, or even higher, inter-annotator agreements, in terms of Kappa and ICC, to those observed between expert-generated annotations.

7 References

1. Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20 (1): 37–46.
2. Fisher, Ronald A. (1954). *Statistical Methods for Research Workers* (Twelfth ed.). Edinburgh: Oliver and Boyd