Nonsense!: Quality Control via Two-Step Reason Selection for Annotating Local Acceptability and Related Attributes in News Editorials

Wonsuk Yang **Seungwon Yoon** **Ada Carpenter** **Jong C. Park**[†]
School of Computing
Korea Advanced Institute of Science and Technology
{derrick0511, swyoon, ada, park}@nlp.kaist.ac.kr

Abstract

Annotation quality control is a critical aspect for building reliable corpora through linguistic annotation. In this study, we present a simple but powerful quality control method using two-step reason selection. We gathered sentential annotations of local acceptability and three related attributes through a crowdsourcing platform. For each attribute, the reason for the choice of the attribute value is selected in a two-step manner. The options given for reason selection were designed to facilitate the detection of a nonsensical reason selection. We assume that a reliable annotation may not contain a nonsensical reason selected for the choice of the attribute value, and an annotation that contains a nonsensical reason is less reliable than the one without such reason. Our method, based solely on this assumption, is found to retain the annotations with remarkable quality out of the entire annotations mixed with those of low quality.

1 Introduction

Crowdsourcing has recently enabled the collection of large amounts of data quickly. However, crowdsourced annotations require more quality control than traditional in-house annotations. This is because it is difficult to assure the expertise of the participating crowdsource workers in the linguistic annotation.

To ensure the quality of the annotations obtained through crowdsourcing, a frequently utilized approach is to request several workers annotate the same data, aggregate the annotations, and infer the ground truth based on the assumption that the consensus of the workers would lead to the ground truth (Lease, 2011). Various statistical methods have also been proposed to filter out the mistakes or (spamming) random responses of the crowdsource workers (Liu et al., 2012; Hovy et al., 2013; Nguyen et al., 2017). However, the way to filter out the mistakes or the random responses through statistical means is difficult to utilize for a fundamentally subjective annotation task. This is because statistical methods make it hard to differentiate the annotator's mistakes from their subjective annotations, as the individual subjectivity of views on different topics can make the gathered data apparently random.

Local acceptability, which is for how much the given text is rationally worthy of being believed to be true (Wachsmuth et al., 2017; Yang et al., 2019), is a fundamentally subjective measure. It is thus hard to utilize only statistical methods for the annotation quality control. Yet, local acceptability is also understood as a strong factor influencing the argumentation quality in writing, which is of great importance for computer-assisted writing. In this work, we present a simple but powerful method for annotation quality control of local acceptability using two-step reason selection, where the options given for reason selection were designed to facilitate the detection of a nonsensical reason selection. In order to show the effectiveness of the method, we performed the sentential annotation of local acceptability and three possibly related attributes through crowdsourcing.

Using our method, we were able to filter out the annotation results with Krippendorff's alpha of 0.3 for local acceptability, which is comparable to our sentence-level in-house annotation of local acceptability (Yang et al., 2019) and the document-level annotation of local acceptability by Wachsmuth et al. (2017). The overall, non-filtered annotation results were with the alpha of 0.01. Our analysis on argumentation strategy indicates that the filtered annotation results match good linguistic intuition whereas those non-filtered do not. The filtered annotation results also show distinctive cor-

[†] Corresponding author

relation patterns whereas those non-filtered do not.

The contributions of this paper are as follows. (1) We present a new method of using reason selection for the quality control of the annotation of local acceptability and three (possibly related) attributes. (2) We provide a statistical analysis of the annotation results and detailed statistics on the filtering and show that our method has a good potential for future research into annotation quality control. (3) We make all the related data and the code for filtering publicly available.

To the best of our knowledge, our method is the first in utilizing nonsensical reason selection for the quality control of the linguistic annotation.

2 Related Work

2.1 Argument mining and the annotation of the reasons

Hasan and Ng (2014) annotated the reasons for ideological debate posts and built the classifiers for such reasons. They set out four target domains, and categorized possible reasons into several groups for each of the two stances of support and opposition. Habernal and Gurevych (2016) annotated the convincingness of arguments through crowdsourcing, with a specified reason for the annotated convincingness. They presented a decision tree that restricts the scope of the possible reasons and enables a structured analysis of the reasons selected by workers. Ding et al. (2018) annotated the reasons for an event being affective, based on seven categories of common human needs. They formalized a classification task for such seven categories by assigning a gold label on the reason for each event. They also showed that baseline classifiers could achieve moderate performance on the task. On the other hand, we focus on local acceptability of arguments with restricted options for reason selection. We choose the reasons not only for a deeper understanding of the reasons for local acceptability judgment itself, but also for an automated identification of the random responses by the workers.

2.2 Quality control of crowdsourced annotation

Raykar et al. (2010) proposed a nonlinear statistical method to infer the actual, hidden label among the multiple and possibly noisy labels provided by multiple annotators, and showed that their method is superior to that of majority voting. Liu et al. (2012) addressed the problem of aggregating the labels from unreliable annotators via crowdsourcing as a standard inference problem in graphical models and showed that the belief propagation model and the mean field model can be effective. Hovy et al. (2013) proposed MACE (Multi-Annotator Competence Estimation) system that can identify the adversary (spamming) annotators utilizing a latent variable that can encode the information of whether or not an annotator is spamming and when the annotator is spamming. Nguyen et al. (2017) utilized Hidden Markov Model to aggregate the sequential labels obtained through a crowdsourcing platform,

	Score	Description
Local Acceptability	7	I **strongly accept** the information given by the sentence to be true. I have sound and cogent arguments to justify my acceptance.
	6	I **accept** the information given by the sentence to be true. I have some arguments to justify my acceptance.
	5	I **weakly accept** the information given by the sentence to be true. I do not have arguments justifying my acceptance. Still, I will accept it rather than reject it.
	4	It is **hard to judge** whether I should accept or reject the information given by the sentence to be true.
	3	I **weakly reject** the information given by the sentence to be true. I do not have arguments for the rejection. Still, I will reject it rather than accept it.
	2	I **reject** the information given by the sentence to be true, and I have arguments for the rejection.
	1	I **strongly reject** the information given by the sentence to be true. I have sound and cogent arguments for the rejection.
Knowledge Awareness	3	I **already knew** the information before I read this document.
	2	I **did not know** the information before I read this document, **but came to know** it by reading the previous sentences in this document.
	1	I **did not know** the information.
Verifiability	5	I can verify it **using my knowledge**. I do not need to google it to verify.
	4	I can verify it by **short-time googling**.
	3	I can verify it by **long-time googling**. I could verify it using deduction if I google it for some time for deeper understanding.
	2	I might find an **off-line way** to verify it, but it will be very hard.
	1	There is **no way to verify** it.
Disputability	4	Whether or not it is reasonable to accept the information given by the sentence as true, it is **highly disputable**.
	3	Whether ... true, it is **disputable**.
	2	Whether ... true, it is **weakly disputable**.
	1	Whether true, it is **not disputable**.

Table 1: Description of Local Acceptability and related attributes.

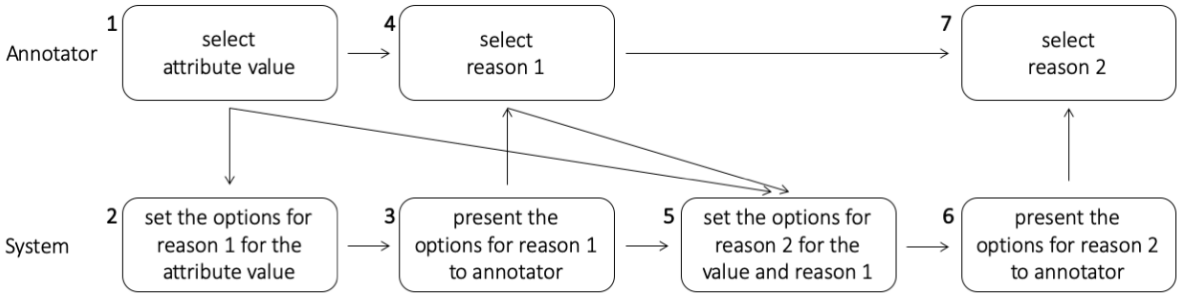

Figure 1: The process of the annotation for an attribute value. For a sentence, the annotator iterates the process four times for the four attributes to annotate. The number at the upper left side of each box indicates the order of the process. The arrow indicates the passing of the information.

and showed that their method works satisfactorily on the sequential labeling tasks of Named Entity Recognition and Information Extraction.

Guillaume et al. (2016) presented ZombiLingo, which is a 'Game with a Purpose' (GWAP) for dependency syntax annotation. Even though the annotation of dependency syntax is expected to require moderate linguistic expertise, they gathered more than 100,000 annotations in 9 months for French dependency syntax with 0.93 precision by decomposing the task into sub-tasks and by including the annotator training as a part of the game. Abad et al. (2017) presented a method for crowdsourced annotation, in which an automatic classifier is used for self-training. For some data with which the classifier shows high confidence for the predicted label, their self-training system guides the annotators to the (highly likely) correct labels predicted by the classifier. For the task of assessing the quality of the machine translated results, Mathur et al. (2018) estimated the reliability of the workers by presenting multiple translations as an assignment. For the translations in an assignment, some were already scored by human experts, and some were presented twice in the assignment. They assessed the reliability of each annotator by seeing how close their annotations were to the expert-annotated results, and whether they annotated the same translations with the same score.

In this study, unlike the previous researches, we use the reason selection method for an automated identification of the random responses by the workers. To the best of our knowledge, there has been no method utilizing reason selection for automated identification, even though incoherence between the selected reason and the attribute value can be a strong indicator for random responses. Our experiment shows that the reason selection method is simple, but very powerful.

3 Data

We used the Webis-Editorials-16 corpus provided by Al-Khatib et al. (2016). For 300 news editorials, they annotated discourse units in each news editorial with the labels that are related to an argumentation strategy, such as (1) *Common Ground*, (2) *Assumption*, (3) *Testimony*, (4) *Statistics*, (5) *Anecdote*, and (6) *Other*. We randomly selected 105 news editorials among the original 300 news editorials in the corpus and annotated local acceptability and the related attributes for each sentence in the 105 news editorials. The news editorials are somewhat out-dated, but we have chosen to use the data with two reasons: (1) Their annotation of argumentation strategy has high and reliabile quality. (2) Our previous sentential annotation of local acceptability (Yang et al., 2019) is based on the news editorial corpus.

4 Annotation

4.1 Definition

Local acceptability is defined, for the premises of an argument, as "rationally worthy of being believed to be true" (Wachsmuth et al., 2017). A premise gives the reason for justifying (or refuting) a claim, where the claim is a possibly controversial statement and the central argument component. A claim and one or more premises compose an argument (Stab and Gurevych, 2017). Following the definition of local acceptability by Wachsmuth et al. (2017), we define the local acceptability *of a sentence*, based on the truth-value of the sentence defined in the *truth-conditional*

Attribute Value	Reason 1	Reason 2
I did not know the information	I am not familiar with the topic	I am an expert on the subject
I did not know the information before, but came to know it by reading the previous sentences	It is not directly stated so far, but I can infer it from the previous sentences	The same information was stated in one of the previous sentences
I can verify it using my knowledge	I can logically verify the information	It is a transition sentence that is meaningless
strong accept	It is not a factual information, but I agree with the statement	It is a fact

Table 2: Examples for nonsensical reason

Attribute Value	Reason 1	Reason 2
weakly disputable	Some may think different on some details	It is a famous conflict
I already knew the information before I read this document	I know it from my personal experience	I am not familiar with the topic
I can verify it by short-time googling	I would find a direct reference to verify it	I can interview the people related to it
accept	It is a subjective statement, but I agree with the statement	It can easily be verified

Table 3: Examples for unnatural reason

theory (Lewis, 1970), as: A sentence is locally acceptable if the truth-value of the sentence is rationally worthy of being believed to be true (Yang et al., 2019).

4.2 Attributes

Table 1 shows the rubrics for local acceptability and the other three attributes that we annotated through crowdsourcing. We introduced the three attributes (and the rubrics) in Yang et al. (2019) for a more in-depth understanding of local acceptability, focusing on the aspects of computational journalism (Cheruiyot and Ferrer-Conill, 2018; Aharoni and Tenenboim-Weinblatt, 2019).

Local Acceptability (LA) is an indicator of whether the truth-value of the sentence is rationally worthy of being believed to be true. We divide each of the *accept* and *reject* judgments into three subcategories. For *accept*, we divided it into *strong accept*, *accept*, and *weak accept*, and the same goes for *reject*. We did so as the judgment on the truth-value can be unclear (Hamblin, 1970).

Knowledge Awareness (KA) is an indicator of whether or not an annotator already knew the information given by the sentence before reading a document.

Verifiability (V) is an indicator of how easy it is to verify the information given by the sentence. We include this attribute because we anticipate that the information recognized as easily verifiable by a reader would be more likely to be accepted by the reader.

Disputability (D) is an indicator of how controversial the information given by the sentence is. We anticipate that the personal acceptance of information by a reader would be different from the reader's expectation of the public acceptance of the information by others.

5 Method

5.1 Two-step reasons for each attribute

For each attribute of a sentence, after the annotator chose its value according to the rubrics, they were asked also to choose the reasons for the value, in a two-step manner. At the first step, the annotator was asked to select one of the possible options as a reason (reason 1). Then, they were asked to select another reason (reason 2), where the options for the second step were determined by the attribute value and the reason selected in the first step. Figure 1 illustrates the process of the attribute value and reason selection.

The options for a reason were designed manually, based on our previous sentence-level in-house annotation result (Yang et al., 2019), mainly because the automatic categorization of reasons is a fundamentally challenging problem (Skorupski, 2002). The in-house annotation was performed by three of the authors and seven undergraduate students with native competence in English, where three of them were student journalists responsible for the school newspaper. The in-house annotation was for the same attributes but the students wrote down the reasons by themselves without any given

Attribute	Attribute Value	Reason 1	Reason 2	label
Local Acceptability	strong accept	It is a subjective statement, but I agree with the statement	It is a fact	Nonsensical
Knowledge Awareness	I did not know the information	I am not familiar with the topic	I am an expert on the subject	Nonsensical
Verifiability	I might find an off-line way to verify it, but it will be very hard	I could visit related places to obtain the information	I can interview the people related to it	Natural
Disputability	weakly disputable	Some may think different on some details	It is a famous conflict	Unnatural

Table 4: An example for the two-step reason selection for a sentence

option. The in-house annotation gathered 4,623 sentential annotations for the 105 news editorials with specific reasons written without any given option. One of the authors thoroughly read all of the 4,623 sentential annotations, and two other authors inspected part of the annotations. Based on the in-house annotation result, three of the authors discussed and designed the options so that they can cover all the responses on the in-house annotation result. Yet, anticipating that there may be a case where the options for the reasons were not enough for a sentence, we allowed the workers to choose *n/a* at each of the three steps to select (1) attribute value, (2) reason 1, and (3) reason 2. When a worker chose *n/a* for an attribute value, we asked the annotator to select the reasons in a two-step manner, and if a worker chose *n/a* for either reason 1 or reason 2, we did not ask the worker for further details. We expect that the selections of *n/a* would enable us to estimate the range of sentences that the option design may not cover.

5.2 Nonsensical/Unnatural Reason

For each attribute, we intentionally added nonsensical/unnatural options for a reason at the second step. Those were designed to identify the mistakes or the (spamming) random responses by the workers. We deemed the selection of the nonsensical/unnatural reason as an indicator for the unreliability of the annotation. The definition of the nonsensical/unnatural reason is as follows.

Nonsensical reason is the reason that a reliable annotator would not select for the annotation and is the reason that works as an indicator for a (spamming) random response. For example, if the reason 'I am not familiar with the topic' for knowledge awareness is selected with the reason 'I am an expert on the subject,' we find that it does not make any sense. Table 2 shows examples for nonsensical reason.

Unnatural reason is the reason that a reliable annotator is less likely select than the other given options, but it is the reason behind which the annotator might have a sophisticated rationale. For example, if an attribute value of *weakly disputable* for disputability is selected with the reason of 'It is a famous conflict,' it makes some sense, but it is deemed as unnatural. This is because if a sentence is about a famous conflict, the attribute values of *disputable* or *highly disputable* would have been more suitable. Table 3 shows examples for unnatural reason.

For each of the triples of (attribute value, reason 1, reason 2) that can be selected by the worker, we assigned one of the following labels: (1) natural, (2) unnatural, and (3) nonsensical. The assignment was also conducted manually by three of the authors. The reader is referred to Appendix A for the entire set of options and the corresponding labels for the two-step reason selection. Table 4 shows an example of the possible selection for a sentence. It should be noted that a sentential annotation can contain more than one nonsensical/unnatural reason. If an annotator chose the options in Table 4 for a sentence, the sentential annotation contains two nonsensical reasons and an unnatural reason.

5.3 Unreliability Score

For sentential annotation, we compute the unreliability score of the annotation as follows:

$$u(x) = 0.5 \times N_u(x) + N_n(x) \quad (1)$$

where x is a sentential annotation, $u(x)$ is the unreliability score of x, $N_u(x)$ is the number of the unnatural reasons over the four attributes within x, and $N_n(x)$ is the number of nonsensical reasons within x. For example, $u(x)$ is 2.5 for the selection in Table 4. Regarding the coefficient 0.5, another value can also be chosen in its place. The unreliability score can be defined as a function that takes $N_u(x)$ and $N_n(x)$ as arguments. In this paper, we are using the linear function with an ad-hoc coefficient of 0.5 to demonstrate the effectiveness of our method as we believe that it is the simplest form of the equation.

For a worker, we compute his or her unreliability score as follows:

$$U(w) = Average(u(x)) \quad (2)$$

for x in $X(w)$ where w is a worker, $U(w)$ is the unreliability score of w, $X(w)$ is the set of annotations annotated by w (possibly over multiple news editorials), and $Average(u(x))$ for x in $X(w)$ indicates the average of the unreliability scores of all the annotations conducted by the worker w.

6 Experiment

6.1 Setting

In this study, the annotators were asked to proceed with the annotation in one direction. In other words, they were asked to read a sentence, to annotate the sentence, then to read the next sentence, and so on. Therefore, before proceeding with the annotation, we built an annotation tool in order to provide an environment that helps unidirectional reading. The annotation tool shows only a single sentence at a time on the screen, and the annotator could proceed to the next sentence by a keystroke. We allowed the annotators to go back to the previous sentence, and modify the previous annotation. The tool also recorded the time of each keystroke.

We conducted the annotation through the Amazon Mechanical Turk (AMT) crowdsourcing platform. In each assignment, the workers were presented with the annotation guideline and URL for a news editorial. For each sentence, workers were asked to select an attribute value for each attribute and to select the reasons for the attribute value in a two-step manner. The workers were asked to proceed to the annotation using the annotation tool. Although we used the corpus of argument strategies as the basis for our annotation, we provided the annotators only with plain sentences, without any extra information hinting at the argument strategies. Moreover, we paid attention not to inform the workers of the fact that we could assess the annotation quality based on the nonsensical/unnatural reason selection during the entire communication with the workers. We also did not reject any worker for the nonsensical/unnatural reason selection.

We made an additional request for the annotations from annotators who completed them too quickly, for instance in less than 20 seconds for all the 12 values for a sentence. If the annotators did not agree to re-do the annotation for the second time, we rejected the first annotation. We made this additional request on 29% of the initially requested assignments. Among these, 90% were re-annotated, and 10% were rejected. For the re-requested annotations, we dismissed the first annotations and took only the re-annotated results for our experiment. As a result, counting only the approved (and not rejected) annotations, 218 workers (with the lifetime approval rate $> 95\%$) participated in our annotation, over about two weeks including the time for the re-request. One annotator was allowed to annotate multiple news editorials, but not the same news editorials multiple times.

Using our method, we ranked the workers based on his or her unreliability score, and divided the workers into three groups as follows: (1) the group of top 1/3 workers with the *lowest* unreliability score per worker, (2) the group of bottom 1/3 workers with the *highest* unreliability score per worker, and (3) the group of the other workers. Among the 218 workers who participated in our annotation, the size of each group was (1) 72, (2) 73, and (3) 73, respectively, as 218 is not a multiple of three.

Then, we grouped the annotations conducted by the annotators in each of the (1)-(3) groups and named the groups of annotations as (1) *good*, (2) *bad*, and (3) *moderate* annotations, respectively. Among the *good* annotations, we filtered out some of the annotations again with the lowest unreliability score per annotation so that the number of retaining annotations in the group would be 10% of the total annotations gathered from our experiment, and named the group as *good squared* annotations. In short, the *good*, *bad*, and *moderate* annotations were filtered based on the unreliability of the annotators, and the *good squared* annotations were those filtered again from the *good* annotations based on the unreliability score of each annotation. It should be noted that the four groups are those of annotations, not those of workers.

6.2 Results

6.2.1 Statistics

We annotated 3,591 sentences in 105 news editorials and gathered 17,414 sentence level annotations. For each of the documents, the average of the number of annotators per news editorial was 4.85. We initially set the number of annotators as five annotators per news editorial, and 3% of the assignments were rejected. For the unrelia-

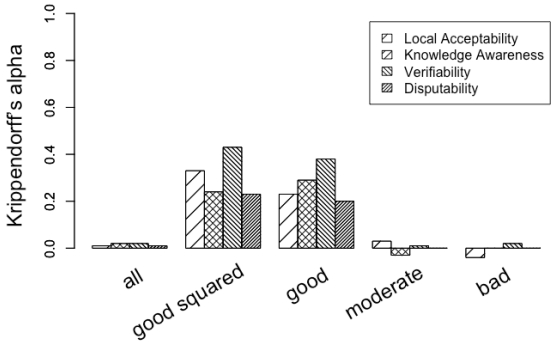

Figure 2: Inter-annotator agreement

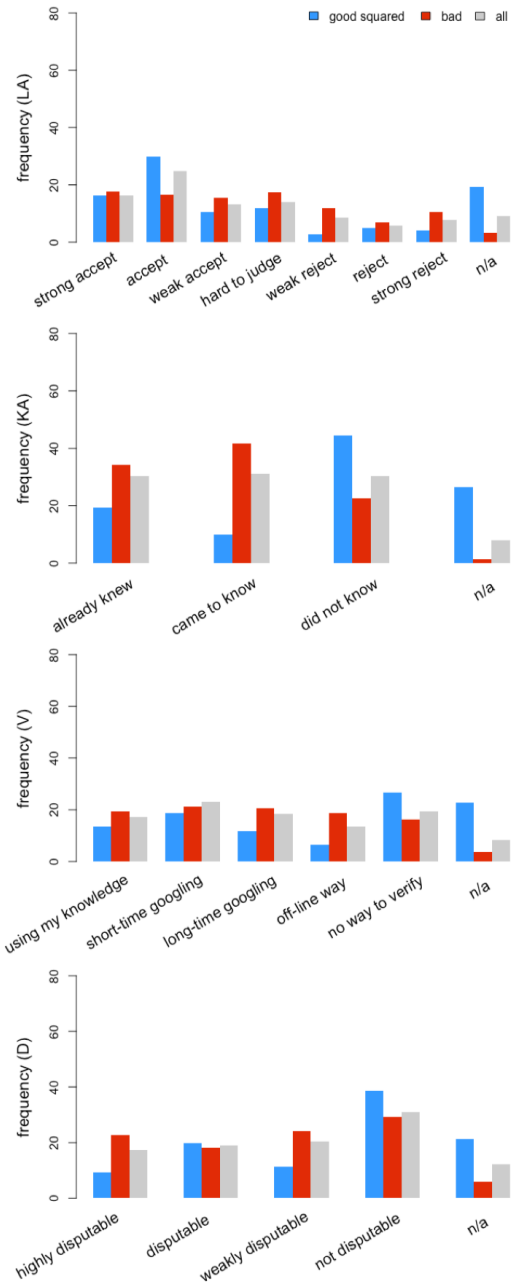

Figure 3: Relative occurrence frequency (unit: %)

bility score of annotation, the average of the unreliability score, over the annotations in the *good squared*, *good*, *moderate*, *bad* annotations, is zero, 0.05, 0.48, and 0.84, respectively. The number of annotations within the *good squared*, *good*, *moderate*, and *bad* group is 1,741 (10%), 3,024 (17%), 6,304 (36%), and 8,086 (46%), respectively.

We speculate that the imbalance on the number of annotations for each of the *good*, *moderate*, *bad* annotations might have been due to the difference in the time that took for the annotation per annotator. The average time that took for a news editorial [1] was 54 minutes per assignment (per document, *i.e.*, per news editorial) for overall annotators. For the *good*, *moderate*, and *bad* annotations, it was 122, 66, and 19 minutes, respectively. For the *good squared* annotations, it was the same as the *good* annotations, as the filtering from the *good* annotations to *good squared* annotations is conducted only on an annotation-level filtering within an assignment, not affecting the calculation. As the workers could be assigned to another news editorial until the study ends, the workers who completed an assignment faster could be given another assignment, hence the different number of annotations.

6.2.2 Inter-annotator Agreement

In order to check for the effectiveness of the quality control method using the two-step reason selection, for each of the four groups, we retained only the annotations by the annotators in the group and measured the inter-annotator agreement (IAA) of the annotations that were retained, dismissing all the other annotations in the other groups. For example, for the *good squared* annotations, we retained only the 1,741 *good squared* annotations among the 17,414 crowdsourced annotations, dismissing all the other 15,673 (=17,414−1,741) annotations and measured the IAA for the 1,741 annotations that were retained. The IAA for *good*, *moderate*, and *bad* annotations was calculated in the same way. Figure 1 shows the inter-annotator agreement for each group. We measured the IAA based on the Krippendorffs alpha, as the attribute values are ordinal. For local acceptability, alpha was 0.33 for *good squared* annotations and 0.23 for *good* annotations. On the other hand, for all of *moderate*, *bad*, and *all* annotations, the IAA for each attribute was less than 0.03, which is very

[1] We dismissed 18 outliers taking time longer than 10 hours for annotating a news editorial.

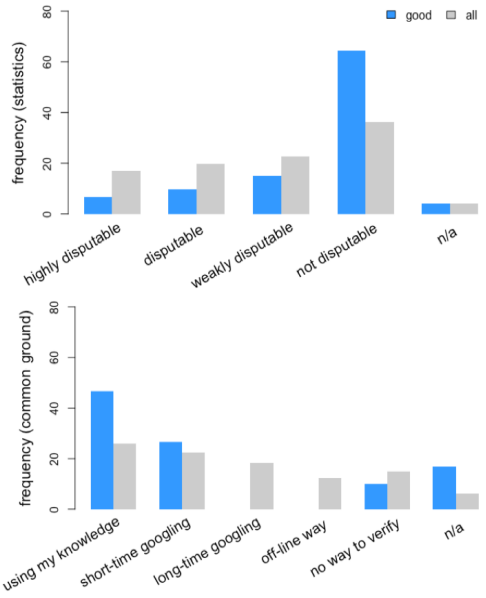

Figure 4: Relative occurrence frequency (unit: %)

low. The alphas of *good squared* annotations and *good* annotations are comparable to those of in-house annotations by Yang et al. (2019), which is the basis for the design of the reason options (see Section 5.1), showed IAA of 0.24 alpha. Wachsmuth et al. (2017) reported that the IAA of the document-level local acceptability annotations was measured between the most agreeing annotator pair among the three annotators and that the result was 0.46. Besides, seven experts who participated in their pilot study are found to show the IAA lower than 0.22 on pilot data, which is comparable to the alpha of our *good squared* and *good* annotations.

We also note that the sample size (the number of annotations) of each of the *good squared*, *good*, *moderate*, and *bad* annotations is different. To make sure that the IAA of *good squared* and *good* annotations is not the result of a different sample size, we randomly sampled 10%, 20%, ..., and 90% of all the annotations. In other words, we adopted random sampling as a baseline method to compare with our proposed method of the two-step reason selection, and checked the effect of the sample size on IAA. For each of the groups of randomly selected annotations with a different sample size, the IAA was less than 0.03 alpha for all the attributes. This confirms that the IAA of *good squared* and *good* annotations is not due to a different sample size of the annotations.

6.2.3 Occurrence Frequency

Figure 2 shows the occurrence frequency of the attribute values for each attribute. It is notable that for local acceptability, the imbalance of the data was quite significant. The occurrence frequencies of *accept* and *strong accept* were much higher than those of the other labels. The *bad* annotations contain fewer *n/a*'s than *all* annotations, or the entire crowdsourced dataset. We find that the *bad* annotations have a more balanced distribution of the attribute values selected for each attribute. We believe that this is an indicator for more randomized selections in the *bad* annotations.

We also looked into the relative frequency of the attribute values for the sentences that contain a discourse unit with a strategy label. We report two cases of *Common Ground* and *Statistics* for the attributes of verifiability and disputability, respectively. Figure 4 shows the relative frequency of the attribute values for the two strategy labels. For *Common Ground*, the distribution of the attribute values for disputability of *good* annotations shows a higher bias to *not disputable* than those of *all* annotations. For *Statistics*, the distribution for verifiability for *good* annotations shows a higher bias to *using my knowledge* than those for *all* annotations. We believe that these biases match good linguistic intuition and that the higher biases of *good* annotations than *all* show the effectiveness of our method.

6.2.4 Correlation Analysis

In Yang et al. (2019), we report that our in-house annotation results show a positive correlation between local acceptability and verifiability, and a negative correlation between local acceptability and disputability. Moreover, knowledge awareness and verifiability show a positive correlation. We see that the correlation pattern is an expected result, based on the definitions of the attributes. Figure 5 shows the Pearson's Correlation Coefficient (PC) between the four attributes for our annotation results using the two-step reason selection. *Good squared* and *good* annotations show the distinctive correlation pattern that is similar to that reported in Yang et al. (2019). *Moderate* and *all* annotations on the other hand show an opaque correlation pattern, and *bad* annotations show no correlation pattern except for a small negative correlation between local acceptability and disputability. We believe that the correlation patterns of the *good squared* and *good* annotations

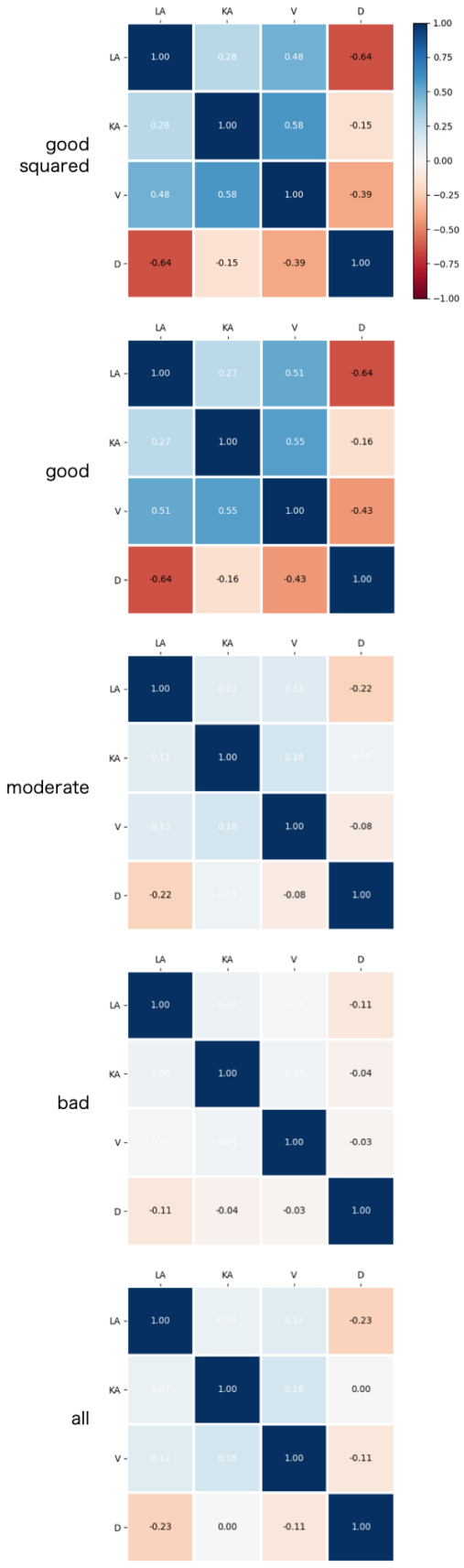

Figure 5: Pearson's Correlation Coefficient between local acceptability and the related attributes.

match good linguistic intuition, and that those of the other annotation groups do not. We also believe that the difference in the correlation pattern between the annotation groups shows the effectiveness of our quality control method using the two-step reason selection.

7 Conclusion and Future Work

In this study, we presented a simple but powerful method of quality control for annotating local acceptability and the related attributes using two-step reason selection. We designed a system in which a nonsensical reason can be selected within a sentential annotation and regarded it as an indicator for the unreliability of the annotation. The analysis of inter-annotator agreement shows that the method is effective at retaining only the annotations with remarkable quality. We conducted our annotation through crowdsourcing *without* the assurance of the expertise of the participating workers in the linguistic annotation, and *without* a separate training phase. Yet, the *good squared* and *good* annotations identified by our method show comparable quality to our previous in-house annotation results which were performed *with* the required expertise and *with* the training for linguistic annotation (Yang et al., 2019). Therefore, we believe that our results show a great potential for the crowdsourced annotation, especially when it is followed by a working quality control method.

One limitation of this study is that the annotators should not know about our use of the nonsensical reason selection as the indicator for unreliability during the annotation. As such, we paid attention not to inform the workers of this and did not reject workers for their selections of the nonsensical reasons. Therefore, we leave it as future work to develop a method that is effective even when workers are aware of quality control methods that employ nonsensical reasons, possibly by making the nonsensical reasons change automatically between annotations.

Acknowledgements

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582-002, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Azad Abad, Moin Nabi, and Alessandro Moschitti. 2017. Self-crowdsourcing training for relation extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 518–523.

Tali Aharoni and Keren Tenenboim-Weinblatt. 2019. Unpacking journalists'(dis) trust: Expressions of suspicion in the narratives of journalists covering the israeli-palestinian conflict. *The International Journal of Press/Politics*.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

David Cheruiyot and Raul Ferrer-Conill. 2018. "fact-checking africa" epistemologies, data and the expansion of journalistic discourse. *Digital Journalism*, 6(8):964–975.

Haibo Ding, Tianyu Jiang, and Ellen Riloff. 2018. Why is an event affective? classifying affective events based on human needs. In *Workshops at the 32nd AAAI Conference on Artificial Intelligence*.

Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3041–3052.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

Charles L. Hamblin. 1970. Fallacies. *Londres, Methuen*.

Kazi S. Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 751–762.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. In *Workshops at the 35th AAAI Conference on Artificial Intelligence*.

David Lewis. 1970. General semantics. *Synthese*, 22(1):18–67.

Qiang Liu, Jian Peng, and Alexander Ihler. 2012. Variational inference for crowdsourcing. In *Proceedings of the 25th International Conference on Neural Information Processing Systems: Volume 1*, pages 692–700.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2018. Towards efficient machine translation evaluation by modelling annotators. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 77–82.

An T. Nguyen, Byron C. Wallace, Junyi J. Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 299–309.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo H. Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

John Skorupski. 2002. The ontology of reasons. *Topoi*, 21(1):113–124.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim A. Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Wonsuk Yang, Jung-Ho Kim, Seungwon Yoon, Chae-Hun Park, and Jong C. Park. 2019. A corpus of sentence-level annotations of local acceptability with reasons. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*.