# Does My Rebuttal Matter? Insights from a Major NLP Conference

**Yang Gao***, **Steffen Eger***,
**Ilia Kuznetsov, Iryna Gurevych**
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

**Yusuke Miyao**
Department of Computer Science
Graduate School of Information
Science and Technology
University of Tokyo
yusuke@is.s.u-tokyo.ac.jp

## Abstract

Peer review is a core element of the scientific process, particularly in conference-centered fields such as ML and NLP. However, only few studies have evaluated its properties empirically. Aiming to fill this gap, we present a corpus that contains over 4k reviews and 1.2k author responses from ACL-2018. We quantitatively and qualitatively assess the corpus. This includes a pilot study on paper weaknesses given by reviewers and on quality of author responses. We then focus on the role of the rebuttal phase, and propose a novel task to predict after-rebuttal (i.e., final) scores from initial reviews and author responses. Although author responses do have a marginal (and statistically significant) influence on the final scores, especially for borderline papers, our results suggest that a reviewer's final score is largely determined by her initial score and the distance to the other reviewers' initial scores. In this context, we discuss the *conformity bias* inherent to peer reviewing, a bias that has largely been overlooked in previous research. We hope our analyses will help better assess the usefulness of the rebuttal phase in NLP conferences.

## 1 Introduction

Peer review is a widely adopted quality control mechanism in which the value of scientific work is assessed by several reviewers with a similar level of competence. Although peer review has been at the core of the scientific process for at least 200 years (Birukou et al., 2011), it is also a subject of debate: for instance, it has been found that peer reviewing can hardly recognize prospectively well-cited papers or major flaws (Ragone et al., 2013). Further, Langford and Guzdial (2015) observed substantial disagreement between two sets of reviews on the same set of submissions for the prestigious Conference on Neural Information Processing Systems (NeurIPS) 2014.

The *rebuttal* phase plays an important role in peer reviewing especially in top-tier conferences in Natural Language Processing (NLP). It allows authors to provide *responses* to address the criticisms and questions raised in the reviews and to defend their work. Although there is evidence that reviewers do update their evaluations after the rebuttal phase[1], it remains unclear what causes them to do so, and especially, whether they react to the author responses per se, or rather adjust to the opinions of their co-reviewers ("*peer pressure*").

In order to obtain further insights into the reviewing process, especially regarding the role of the rebuttal phase in peer reviewing, in this work we present and analyze a review corpus of *the 56th Annual Meeting of the Association for Computational Linguistics (ACL-2018)*. Every reviewer/author was asked whether she consented to freely using her review/author-response for research purposes and publishing the data under an appropriate open-source license within at earliest 2 years from the acceptance deadline (see supplementary material for the original consent agreement). 85% reviewers and 31% authors have consented to sharing their data. The corpus comprises over 4k reviews (including review texts and scores) and 1.2k author responses. Uniquely, the corpus includes both before- and after-rebuttal reviews for both accepted and rejected papers, making it a highly valuable resource for the community to study the role of the rebuttal phase. The corpus as well as our source code and annotations are publicly avail-

* Equal contribution.

[1] For example, see discussions at https://naacl2018.wordpress.com/2018/02/04/analysis-of-long-paper-reviews/ and https://acl2017.wordpress.com/2017/03/27/author-response-does-it-help/.



*Proceedings of NAACL-HLT 2019*, pages 1274–1290
Minneapolis, Minnesota, June 2 - June 7, 2019. 

able at https://github.com/UKPLab/naacl2019-does-my-rebuttal-matter.

Our contributions are threefold. First, in §3, we assess the corpus both *quantitatively* (e.g., correlating Overall Score with aspect scores such as Originality and Readability) and *qualitatively* (e.g., identifying key terms that differentiate "good" from "bad" author responses, annotating paper weaknesses given by reviewers, and rating the quality of individual author responses). Second, in §4, we develop a model to predict whether a reviewer will increase/decrease/keep her initial scores after the rebuttal. We do so in order to analyze and disentangle the sources of review updates during the rebuttal stage. We find that factoring in the author responses only marginally (but statistically significantly) improves the classification performance, and the score update decision is largely determined by the scores of peer reviewers. Third, in §5, we discuss multiple types of biases in the score update process, some of which potentially undermine the 'crowd-wisdom' of peer reviewing.

## 2 Related Work

Several sources provide review and author response data. Since 2013, the NeurIPS main conference publishes the reviews of accepted papers and their author responses. However, these reviews only include the review texts for after-rebuttal reviews. Also, reviews of rejected papers and author responses are not published. Some Machine Learning and NLP conferences, for instance ICLR (International Conference on Learning Representations) and ESWC (Extended Semantic Web Conference), adopt the *open review* model, which allows anyone to access the reviews and author responses. However, most major NLP conferences have not yet adopted the open-review model, and the reviews and author responses in open- and non-open-review venues are likely to be different because people behave differently when their actions are observable (Andreoni and Petrie, 2004).

Kang et al. (2018) provide a corpus of computer science papers from ACL, NeurIPS, CoNLL (The SIGNLL Conference on Computational Natural Language Learning) and ICLR, together with the accept/reject decisions and reviews for a subset of the papers. They suggest several tasks with respective baselines, such as predicting review aspect scores from paper- and review-based features. However, their corpus contains neither before-rebuttal reviews nor author responses, and the size of their review set from NLP conferences (only 275 reviews from ACL-2017 and 39 reviews from CoNLL-2016) is much smaller than ours.

Hua et al. (2019) compile a corpus consisting of 14.2k reviews from major NLP and machine learning conferences. In addition, they annotate 10k argumentative propositions in 400 reviews, and train state-of-the-art proposition segmentation and classification models on the data. But similar to Kang et al. (2018), their corpus does not include before-rebuttal reviews or author responses.

Several publications specifically address the peer reviewing process. Falkenberg and Soranno (2018) investigate what makes a paper review helpful to a journal editor within a specific scientific field. Birukou et al. (2011) and Kovanis et al. (2017) discuss the shortcomings of the review process in general, such as its inability to detect major flaws in papers (Godlee et al., 1998) and its ineffectiveness in selecting papers that will have high citation counts in the future (Ragone et al., 2013). They discuss alternatives to the standard review process such as crowd-based reviewing and review-sharing, i.e., resubmitting a rejected work to another venue along with its past reviews. Ragone et al. (2013) analyze peer reviews across nine anonymized computer science conferences and, among others, identify reviewer biases of multiple types (affiliation, gender, geographical, as well as rating bias: consistently giving higher or lower scores than other reviewers) and propose means for debiasing reviews. However, none of these works quantitatively measures the influence of the rebuttal phase on the final review scores, nor do they provide any corpora facilitating such studies.

Our work is also related to *meta science*, which studies the scientific process in general, i.e., how scientific information is created, verified and distributed (cf. Fortunato et al. (2018)). In this context, our work can be seen as a study on how scientific information is verified.

## 3 Review Corpus

ACL-2018 adopts a reviewing workflow similar to that of other major NLP conferences: after paper assignment, typically three reviewers evaluate a paper independently. After the rebuttal, reviewers can access the author responses and other peer reviews, and discuss their viewpoints. Reviews

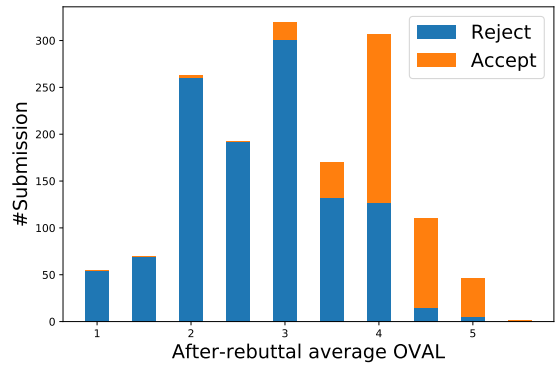


Figure 1: Distribution of accept/reject decisions.

include both *scores* (Overall Score OVAL, Reviewer Confidence CONF, Soundness SND, Substance SBS, Originality ORG, Meaningful Comparison CMP and Readability RDB) and free-text comments. OVAL are integers in $[1, 6]$, while all other scores are integers in $[1, 5]$.

We first provide an overview of our corpus in §3.1, and then present analyses for the reviews and author responses in §3.2 and §3.3, respectively.

### 3.1 Overview of the Corpus

The corpus has three parts: the before-rebuttal reviews (including review texts and scores), the after-rebuttal reviews, and the author responses. The corpus does not contain the submissions, nor the information of the reviewers, e.g., their gender, country, affiliation or seniority level; nevertheless, we perform some analyses on the submissions and the reviewers' information and present the statistics in the supplementary material.

Basic statistics of our corpus are summarized in Table 1. 1542 submissions (1016 long, 526 short) have at least one review opted in. 1538 submissions have at least one before- and one after-rebuttal review opted in. Among the 1542 submissions, 380 submissions (24.6%) were accepted: 255 long, 125 short, and the remaining 1162 were rejected: 761 long, 401 short. The distribution of their accept/reject decisions is illustrated in Fig. 1.

### 3.2 Reviews

**Score Correlation.** In line with Kang et al. (2018), we first assess the impact of individual aspect scores on the overall score by measuring their Pearson correlation, illustrated in Fig. 2. We find that OVAL is most strongly correlated with SND and SBS, followed by ORG and CMP. CONF shows weak positive correlation to RDB: the less readable a paper is, the less confident the reviewers will be. Note that our correlation results are different from those reported by Kang et al. (2018), who report that the OVAL has low Pearson correlation with SND (0.01) and ORG (0.08). While the differences might be caused by a variation in aspect definitions, we believe that our estimate is more reliable as the dataset analyzed in Kang et al. (2018) is substantially smaller than ours.

| Category | Size |
|---|---|
| Before-rebuttal reviews | 3875 (1213 reviewers, 1538 submissions) |
| After-rebuttal reviews | 4054 (1275 reviewers, 1542 submissions) |
| Author responses | 1227 (499 submissions) |

Table 1: Statistics of the ACL-2018 corpus. Some reviewers submitted their reviews after the rebuttal started, hence the size of the after-rebuttal reviews is larger than that of the before-rebuttal reviews.

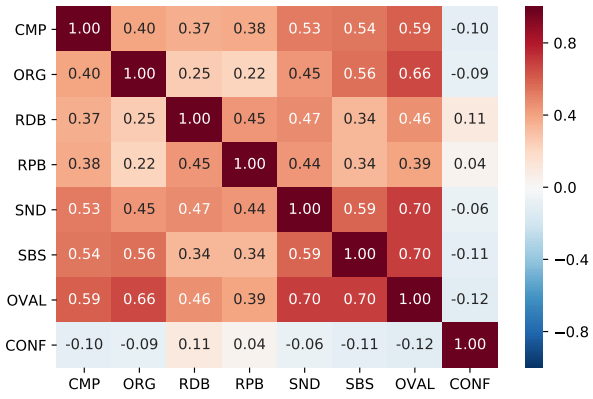


Figure 2: Score correlation matrix.

**Review Texts.** ACL-2018 adopts the novel *argument-based* review template, which asks reviewers to provide *positive* and *negative* arguments for and against the submission, respectively. In addition, reviewers can also list their questions to the authors in the *questions* section of the review template. Most reviewers made good use of the argument-based template: among the 4054 after-rebuttal reviews, 3258 (80.4%) provide positive arguments, 3344 (82.5%) provide negative arguments, and 1627 (40.1%) provide questions. The number and length of arguments/questions are summarized in Table 2.

**Score Changes.** Table 3 shows how many reviews increase (INC), decrease (DEC) or keep (KEEP) their overall scores after rebuttal. For

| Component | Number | Length (token) |
|---|---|---|
| Pos. Arg. | 1.92±1.31 | 22±17 |
| Neg. Arg. | 2.38±1.56 | 56±53 |
| Questions | 0.87±1.36 | 35±31 |

Table 2: Numbers and lengths of different components in each review (mean±standard deviation).

| Type | Num. | #Paper | Acpt.% | $\Delta_{\text{OVAL}}$ |
|---|---|---|---|---|
| INC | 245 | 227 | 49.8 | 2.65 → 3.76 |
| DEC | 248 | 221 | 7.2 | 4.17 → 3.04 |
| KEEP | 3377 | 1119 | 22.8 | 3.13 → 3.13 |
| Total | 3870 | 1538 | 24.7 | 3.17 → 3.17 |

Table 3: Statistics of different types of reviews.

the 227 papers that receive at least one INC review (first row in Table 3), their acceptance rate is 49.8%, much higher than those 221 papers with at least one DEC (7.2%) and those 1119 papers with no score update (22.8%). Hence, the score update has a large impact on the final accept/reject decision. Note that 29 papers receive both INC and DEC reviews, of which five were accepted finally.

Fig. 3 summarizes the OVAL updates. Most reviewers stick to their initial scores after rebuttal. For those who update, the score change usually amounts to just one point in absolute value. However, most updates happen in the borderline area (overall score 3-4) where the score update might influence the overall acceptance decision. We find that the changes in aspect scores occur much less often than the changes in overall scores: only 5% of the reviews have any of the aspect scores updated after rebuttal, and only 1% of the reviews change the confidence value. In these rare cases, aspect score changes are consistent with their OVAL changes, e.g., if the OVAL increases, no aspect score decreases.

**Submission Time.** Fig. 4 illustrates the distribution of the first submission time of reviews. 51.6% reviews were submitted within the last 3 days before the deadline. We also find that the mean submission time of the INC reviews is around 20 hours earlier than that of the DEC reviews, and the difference is statistically significant (p-value 0.009, double-tailed t-test). Moreover, we find that submission time is weakly positively correlated with initial score, which means that reviewers who submit early have slightly lower scores on average, which may explain their tendency to increase their scores later on, given our results in §4.

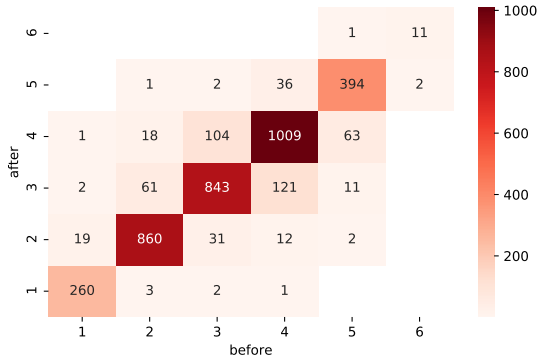


Figure 3: Before vs after rebuttal OVAL.

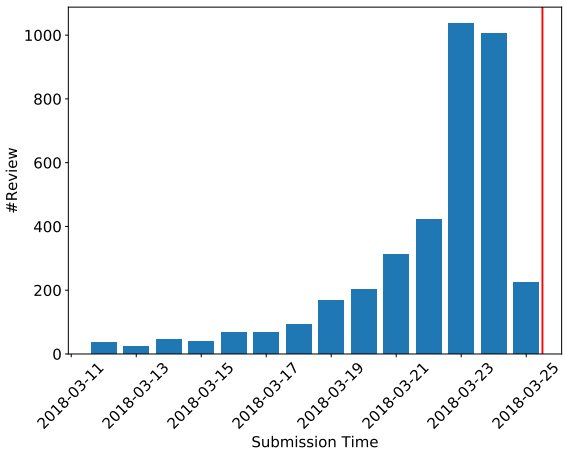


Figure 4: Distribution of review submission time. The review submission deadline (26th March 2018) is marked as the red vertical line towards the right end.

A reason why early submitters have lower scores may be that it takes less time to reject a paper, as the majority of papers is rejected anyway.

**Criticism in Reviews.** To study the most common paper weaknesses identified in reviews, we manually assess about 300 weakness statements from the reviews. Table 4 summarizes the main results, excluding concerns about *Technical weaknesses*. In our sample, most weaknesses refer to *Evaluation & Analysis*, i.e., criticize the lack of: error analysis, ablation tests, significance tests, human evaluations (opposed to indirect measures such as BLEU) and strong baselines as well as insufficient comparisons (either external or internal). Other frequent targets of criticism are *Writ-*

| Eval | Writing | Nov | Data | Motivation |
|---|---|---|---|---|
| 28% | 18% | 8% | 8% | 5% |

Table 4: Frequent weakness types identified in reviews.

| Type | Num. | Length (token) |
|---|---|---|
| iResp | 100 | 373±191 |
| dResp | 80 | 260 ±140 |
| kResp | 1047 | 297±182 |
| Total | 1227 | 300±181 |

Table 5: Statistics of author responses (mean±standard deviation for Length).

*ing* quality, as well as *Data*: e.g., too few datasets being used (only English data or only synthetic data), missing agreement scores for newly labeled datasets, and resources not being publicly available. Reviewers also criticize the lack of *Novelty* and proper *Motivation* of approaches.

### 3.3 Author Responses

We align author responses with their corresponding reviews (if opted in), and term the author responses corresponding to INC, DEC and KEEP reviews as iResp, dResp and kResp, respectively. Table 5 presents an overview on these groups.

To qualitatively compare iResps and dResps, we extract and rank n-grams in both iResps and dResps according to the *log-likelihood ratio* (LLR) statistic (Dunning, 1993), treating iResps and dResps as two corpora[2]. The results are reported in Table 6. We find both iResps and dResps express gratitude and promise revisions in the final versions, but iResps address review questions and criticisms by referring back to certain lines and tables in the original paper while dResps fail to do so. We revisit these differences in §4.2.

| iResps | dResps |
|---|---|
| the final version | thanks for your |
| in line DIGIT | DIGIT reply to |
| in table DIGIT | to question DIGIT |
| in the final | will add more |
| for example the | DIGIT we will |
| in order to | thank the reviewer |
| final version EOS | argument DIGIT reply |
| due to space | due to the |
| for your comments | paper is accepted |
| camera ready version | the revised version |
| DIGIT and DIGIT | we agree that |

Table 6: Top trigrams based on LLR ranking. All digits were replaced by DIGIT. EOS: end of sentence.

To gain further insights, we analyze the *quality* of the author responses to the 300 weakness statements from Table 4. We advertised no formal definition of quality, and assessed a subjective, perceived quality score in a range from 1 (low) to 10 (high). We find that the *weak* author responses (scores 1-3) are substantially shorter than the *strong* ones (scores 8-10): the average token number in weak and strong responses are 53 and 90, respectively. Responses evaluated as weak are less specific and make vague promises (*"Thanks for the suggestion, we will try this in the camera-ready"*), off-topic (addressing different points than those raised by the reviewer), or apologetic (*"the deadline was very close", "our native language is not English"*). Interestingly, with some exceptions (*"We take your review as an example of bad writing"*), the weak responses are usually polite and admit the weaknesses suggested by the reviewers, but they tend not to detail how they would address the weaknesses. Strong responses, in contrast, are specific (referring to specific line numbers in the submission, as well as providing numerical values), detailed, longer, and often do not agree with the criticism, but explain why the reviewer's requirement is hard to meet or beyond the scope of the work.

## 4 After-Rebuttal Score Prediction

To measure the influence of different factors on the score update decisions, we propose and study the *after-rebuttal score prediction* task. Because most score updates after rebuttal do not exceed 1 point (see Fig. 3), we formulate this problem as a classification task. Specifically, given a before-rebuttal review, its corresponding author response and other peer reviews, we try to predict whether the reviewer will increase (INC), decrease (DEC) or keep (KEEP) her overall score after the rebuttal. We avoid predicting the final accept/reject decisions because they are not only based on the final scores (see Fig. 1, where a few low-score papers are accepted while some high-score papers are rejected), but also based on additional factors such as the balance of areas and diversity of papers, which are difficult to measure. The score updating of reviews, in contrast, only depends on the peer reviews and the authors responses.

We choose a classic feature-rich classification model for this task, for two reasons: a) model capacity is lower compared to, e.g., a deep neural network, which is beneficial in our small data scenario, and b) the results are easier to interpret.

[2]We only include n-grams that appear in at least 7 different author responses.

### 4.1 Features

**Score features (`Score`).** We use all peer review scores for a given submission to build an array of score-based features. These include review $i$'s before-rebuttal OVAL (`self_score`), statistics of the other peer reviews' OVAL (denoted by `oth_X`, where `X` can be `max`/`min`/`mean`/`median`/`std`), statistics of all peer reviews' OVAL (`all_X`), and elementary arithmetic operations on the above features (e.g., `oth_mean-self` denotes the mean OVAL of the peer reviews minus review $i$'s before-rebuttal OVAL). CONF are considered in a similar manner. We do not consider aspect scores such as ORG because they yielded no improvements in our preliminary experiments. The full list of features can be found in the supplementary material. We also include features based on the author response texts, as detailed below.

**Length of response (`log_leng`).** We have found that high-quality author responses are usually longer than the low-quality ones (see §3.3). We use the logarithm of the number of tokens in author responses as a feature.

**Review-Response Similarity (`sim`).** Lack of similarity between a review and its response may indicate that the response is "off-topic". To measure similarity, we have trained 300-dimensional skip-gram word embeddings on 5611 papers extracted from the cs.CL (computational and language) and cs.LG (learning) categories of ArXiv which were published between January 1, 2015 and December 31, 2017. We represent reviews and responses by averaging the embeddings of their words, and measure semantic similarity by cosine similarity.[3] We find it important to use word embeddings trained on *CL/LG* domain data: for example, nearest neighbors of "neural" in a model trained on Wikipedia are "axonal", "salience", while on Arxiv its nearest neighbors are "feedforward" and "deep". We find that iResps are more similar to their reviews than dResps and kResps: the average cosine similarity between the reviews and iResps, dResps and kResps are .38, .30 and .29, respectively.

**Specificity (`spec`).** In our human annotation experiments, unspecific responses were typically judged as weak because they did not address specific questions or weaknesses given by reviews. To measure the specificity of author responses, we use a feature-rich sentence-level specificity model by Li and Nenkova (2015) trained on multiple news corpora. The produced scores are in the $[0, 1]$ range, with higher values meaning higher specificity. iResps are slightly more specific than the other responses: the mean specificity scores for iResps, dResps and kResps are .29, .24 and .28, respectively. For each author response, we compute the `spec` scores for all their sentences and use statistics (`max`/`min`/`mean`/`median`/`std`) of the `spec` scores as features. The same strategy is used to build the politeness and convincingness features introduced below.

**Politeness (`plt`).** We employ the sentence-level politeness framework suggested by Danescu-Niculescu-Mizil et al. (2013) to quantify the politeness of the author responses. We have trained a simple bag-of-words based multi-layer perceptron (MLP) model using their Wikipedia and StackExchange data and applied it to the author responses, generating a politeness score in $[-1, 1]$ for each sentence in author responses, where higher scores mean higher politeness. While the mean politeness scores in iResps, dResps and kResps have no marked differences (all around 0.19), the score for the most polite sentence in iResps (.91) is higher than that of dResps (.68) and kResps (.90).

**Convincingness (`cvc`).** To approximate rebuttal convincingness we use the sentence-level convincingness model developed by Simpson and Gurevych (2018), trained on ∼1.2k argument pairs from web debate forums. We normalize all convincingness scores to $[0, 1]$, where larger scores mean higher convincingness. Mean convincingness scores for iResps, dResps and kResps are .60, .49 and .58, respectively.

**Score validation.** Since the `spec`, `plt` and `cvc` models are not trained on review-rebuttal data, we need to perform human evaluations to validate the produced scores. We rank the sentences in author responses in terms of their `spec`, `plt` and `cvc` scores and analyze the top and bottom 10 sentences in each ranking (see the supplementary material). We find that the scores successfully distinguish the most and least specific/polite/convincing sentences. To further validate the scores, for each type of score, we have randomly sampled 15 pairs of sentences from author responses and presented the pairs to 3 ex-

[3] We also used ROUGE (Lin, 2004) to measure the similarity but find the ROUGE scores to be highly correlated with the cosine similarities (Pearson correlation $> 0.9$), so we include only the cosine similarities in our models.

| | spec | plt | cvc |
|---|---|---|---|
| Inter-User | .87 | .87 | .64 |
| User-Score | .93 | .87 | .67 |

Table 7: Percentage of agreement for spec, plt and cvc scores. "User-Score" means the agreement between the aggregated (by majority voting) users' preferences and score-induced preferences.

perienced annotators, asking them to indicate the more specific/polite/convincing sentence in each pair. The agreement is presented in Table 7. The agreement between the users' aggregated preferences and score-induced preferences is quite high for all three types, confirming the validity of the scores. Note that the agreement for cvc is lower than the other two; the reason might be that it is difficult even for humans to judge convincingness of arguments, particularly when evaluated on the sentence level without surrounding context nor the corresponding review. The distribution of the spec, plt and cvc scores for iResps, dResps and kResps is in the supplementary material.

### 4.2 Results and Analyses

We perform experiments on a subset of the corpus which only includes the submissions that have author responses and three or more reviews opted in. We term this subset of the corpus *Submissions with Complete Reviews (Full)*. Training models on submissions with fewer reviews would bias certain features (e.g. all_mean) and thus bias the trained models. Also, we separate out the submissions from the Full set whose before-rebuttal average OVAL are between 3 and 4.5 (note that OVAL are in $[1, 6]$), so as to train and test a model specifically on borderline submissions for which score changes may be decisive for an accept or reject decision. We term this subset *Borderline Submissions (BRD)*. Full includes 791 submissions (80 INC, 60 DEC, 652 KEEP) and BRD includes 590 (69 INC, 48 DEC and 474 KEEP). All results and weights presented in this section are averaged over 5000 repeats of 10-fold cross validation; data entries are randomly shuffled for each repeat.

**Feature Selection.** We filter out features whose information gain is ranked in the bottom 50% of all features on the training set. For highly correlated features in the upper 50% (i.e. Pearson correlation $\geq 0.5$), we filter out all but the one with the highest information gain. Remaining features

| Feature Set | BRD | Full |
|---|---|---|
| spec | .324 | .309 |
| plt | .306 | .310 |
| cvc | .303 | .304 |
| log_leng | .340 | .341 |
| sim | .323 | .302 |
| Score | **.495** | **.526** |
| All but Score | .343 | .336 |
| All | **.522** | **.540** |
| Majority Baseline | .297 | .301 |
| Random Baseline | .258 | .251 |

Table 8: Macro F-1 scores.

are used to train a multinomial logistic regression model (i.e., MLP with no hidden layer and softmax activation function in the output layer). To balance the number of instances for the three classes, on the training set, in each fold of cross-validation we randomly down-sample cases with class KEEP to ensure that the number of KEEP is the same as the sum of INC and DEC. We also tried random forest, decision tree, support vector machines and Gaussian processes as classifiers, but their performances were similar or worse than that of logistic regression.

**Results.** Classification results are presented in Table 8. In addition, we compare to two baselines: the *majority baseline* always picks the majority decision (in our case, KEEP); the *random baseline* selects an action at random. Full results, including precision, recall and F1-scores for each label, can be found in the supplementary material.

We find that score-based features are most effective among all features. However, text-based features are also useful, supported by the observations that: **(i)** models using only text features all significantly (p-value $< 0.01$, double-tailed t-test) outperform the majority and random baseline; and **(ii)** using all features gives the best performance, significantly (p-value $< 0.01$) better than using any feature set alone.

Among the non-Score features, log_leng performs best. But we find it has high correlation with multiple Score features, and hence when all features are used, it is filtered out. The features spec and sim perform much better in BRD than in Full, which suggests that, for borderline papers, more weight is placed on whether the response explicitly addresses the points raised in reviews (similarity) and the specificity of the response.

**Analysis.** To interpret our results, we study the weights of the features in our logistic regression model shown in Tables 9 and 10. We observe the following trends:

- **"Peer pressure" is the most important factor of score change**: in both Full and BRD, features reflecting the gap between own and others' review scores (`oth_mean-self` and `self-oth_min`) have by far the largest weights compared to other feature groups. For example, in Full, the `Score` features have (absolute) weights of 0.4 or higher for the class INC, while all other features are substantially below 0.2. The weights make intuitive sense: e.g., when the mean of the other reviewers' scores is above a reviewer's initial score, she has a strong tendency to increase her own score and not to decrease her own score. Similarly, when a review contains a very convincing sentence, this substantially decreases the probability of a score decrease.

- **To improve the score for a borderline paper, a more convincing, specific and explicit response may be helpful**: in Full, no weight of a text-based feature is above 0.2 for INC; however, in BRD, the weights for `cvc_min`, `spec_median` and `sim` are all above 0.2. This asymmetry of the text-based features across Full and BRD also suggests that reviewers do appear to pay more attention to the author responses in situations where they may matter (e.g., make the difference between accept or reject decisions).

- **An impolite author response may harm the final score**: in both Full and BRD, the weight of `plt_max` is negative for DEC. In addition, in Full a more polite response helps increase the final score (positive weight for INC, close to 0 weight for KEEP). In BRD, in contrast, a more polite response may not increase the score but only keep it unchanged (positive weight for KEEP, close to 0 weight for INC). If we take BRD papers as those for which the author responses really matter, this means that politeness has an asymmetrical effect: it may push a paper below the acceptance threshold, but not above it. Indeed, `plt_max` is the second best text-feature for predicting decrease for BRD papers.

| Feature | INC | DEC | KEEP |
|---|---|---|---|
| `oth_mean-self` | 1.044 | -1.265 | .221 |
| `self-oth_min` | -.378 | .188 | .190 |
| `cvc_max` | .078 | -.271 | .193 |
| `spec_median` | .159 | -.224 | -.065 |
| `plt_max` | .170 | -.174 | .004 |
| `sim` | .019 | .099 | -.119 |
| `spec_max` | .022 | .029 | -.051 |

Table 9: Feature weights in multinomial logistic regression trained on Full.

| Feature | INC | DEC | KEEP |
|---|---|---|---|
| `oth_mean-self` | .855 | -1.026 | .171 |
| `self-oth_min` | -.372 | .191 | .181 |
| `cvc_min` | .224 | -.258 | -.034 |
| `spec_median` | .293 | -.122 | -.171 |
| `sim` | .214 | -.161 | -.053 |
| `cvc_max` | .117 | -.085 | -.033 |
| `plt_max` | .016 | -.192 | .176 |

Table 10: Feature weights in multinomial logistic regression trained on BRD.

## 5 Discussion

The opinion update process we have described in §4.2 is closely related to the work on *opinion dynamics* (DeGroot, 1974; Acemoglu and Ozdaglar, 2011), which studies how human subjects change their opinions as a reaction to those of peers.

The "peer pressure" effect (opinions being updated to mean opinions) is widely observed in opinion formation of human subjects in controlled experiments. Lorenz et al. (2011) find that in simple estimation tasks ("What's the population density of Switzerland?"), human subjects tend to lean towards a consensus once they are exposed to the opinions of others. Similarly, Moussaid et al. (2013) find two dominant effects for simple factual questions: human subjects tend towards the mean opinion and towards the opinions of highly confident individuals. Our experiments also show that the mean opinion plays a very prominent role in peer reviews, but they show no evidence supporting the confidence effect: features based on the confidence scores do not play a significant role in deciding the final scores (see §4.2). We believe this is due to two main differences between peer reviewing and the controlled experiments in the above works: (i) there does not exist a ground-truth score for a submission, while such true answers about factual questions do exist in the controlled experiments; and (ii) participants of the controlled experiments lose money if they give in-

correct answers, but a reviewer loses nothing when she does not adjust to a (self-assessed) expert.

Three types of *biases* have been studied in explanatory models of opinion dynamics in recent years. The first is opposition between members of different groups (e.g., due to *group-identity*) leading to distancing from certain subjects' opinions (Altafini, 2013; Eger, 2016). The second is *homophily*: individuals ignore opinions too different from their own (Deffuant et al., 2000; Hegselmann and Krause, 2002). The third is *conformity* (Buechel et al., 2015), i.e., the desire to conform to a group norm/opinion. Conformity bias can be strong and persist even in the presence of overwhelming evidence that a group opinion is wrong (Asch, 1951). Our observation that reviewers tend to converge to the mean of all reviews (§4.2) suggests that conformity bias also plays a prominent role in peer reviewing. We found no evidence (on an aggregate level) for the other two biases.

To summarize, conformity bias is the main bias we identified in the peer reviewing process. However, conformity bias has a negative effect on crowd-wisdom in estimation tasks (Lorenz et al., 2011), which strengthens confidence of human subjects in the correctness of their converged answer, while the actual correctness of their consensus is often even worse than the mean of multiple independent answers. A simple method to reduce conformity bias is to blind reviewers from each other, only allowing reviewers to update their reviews based on the author responses; the area chair (who can see all reviews for a paper) is then responsible for considering all (possibly conflicting) reviews and making the accept/reject recommendation. We believe that peer reviewing is to a large degree an opinion dynamics process, a neglected insight hitherto, and that lessons from this field should therefore be beneficial for peer reviewing for NLP conferences and beyond.

Finally, concerning the helpfulness of individual review based feature groups, we believe it reflects a weakness of the current rebuttal stage that politeness does matter, because this is *merely* a social aspect unrelated to the quality of the assessed papers. However, we also showed that filling up author responses with "thank you"s is unlikely to increase a reviewer's score for a borderline paper—so at least, authors do not seem to be able to sneak their papers in via social effects.

## 6 Conclusion

We presented a review corpus consisting of over 4k reviews and 1.2k author responses from ACL-2018. To the best of our knowledge, it is the first corpus that includes both before- and after-rebuttal reviews for both accepted and rejected papers in a major NLP conference. We qualitatively and quantitatively analyzed the corpus, including a manual classification of paper weaknesses outlined by reviewers and a quality rating study of the corresponding author responses.

In addition, we proposed a classification model to predict whether a reviewer will increase/decrease/keep her overall score after rebuttal. By analyzing the feature weights in our model, we quantitatively measured the importance of different decision variables for score updates. We found that the gap between a reviewer's initial score and her peers' scores is the main explanatory variable. Rebuttal-related factors like convincingness, specificity and politeness of responses are considerably less important but still have a statistically significant effect, especially for borderline papers.[4] Our findings shed light on the predominant role of the *conformity bias* in peer reviewing (see §5), and we discuss alternative peer review models addressing this bias. We hope our analyses will help the community better understand the strengths and weaknesses of the current peer review workflow, spurring further discussions.

Finally, provided that the rebuttal phase remains a key feature in many peer reviewed conferences, we think that our novel after-rebuttal score change prediction task can be practically beneficial for authors to restructure their author responses and thereby make them more effective.

## Acknowledgements

The authors thank the anonymous reviewers and Dan Jurafsky for constructive comments and helpful remarks. This work has been supported by the ArguAna Project GU 798/20-1 (DFG), the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1), and the DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1). Ilia Kuznetsov has been supported by the FAZIT Foundation.

[4] We believe that they might become more important when further argumentation/text-based features are integrated.

## References


Daron Acemoglu and Asuman Ozdaglar. 2011. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49.

Claudio Altafini. 2013. Consensus problems on networks with antagonistic interactions. *IEEE Trans. Automat. Contr.*, 58(4):935–946.

James Andreoni and Ragan Petrie. 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of public Economics*, 88(7-8):1605–1623.

Solomon E. Asch. 1951. Effects of group pressure upon the modification and distortion of judgments. *Groups, Leadership, and Men*, pages 177–190.

Aliaksandr Birukou, Joseph Wakeling, Claudio Bartolini, Fabio Casati, Maurizio Marchese, Katsiaryna Mirylenka, Nardine Osman, Azzurra Ragone, Carles Sierra, and Aalam Wassef. 2011. Alternatives to peer review: Novel approaches for research evaluation. *Frontiers in Computational Neuroscience*, 5:56.

Berno Buechel, Tim Hellmann, and Stefan Klößner. 2015. Opinion dynamics and wisdom under conformity. *Journal of Economic Dynamics and Control*, 52(C):240–257.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 250–259.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems (ACS)*, 03(01n04):87–98.

Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Steffen Eger. 2016. Opinion dynamics and wisdom under out-group discrimination. *Mathematical Social Sciences*, 80(C):97–107.

Laura J. Falkenberg and Patricia A. Soranno. 2018. Reviewing reviews: An evaluation of peer reviews of journal article submissions. *Limnology and Oceanography Bulletin*, 27(1):1–5.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379).

Fiona Godlee, Catharina R. Gale, and Christopher N. Martyn. 1998. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. *JAMA*, 280(3):237–240.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5:1–24.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, USA, June 2 - 7, 2019, Volume 2 (Short Papers)*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard H. Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1647–1661.

Michail Kovanis, Ludovic Trinquart, Philippe Ravaud, and Raphaël Porcher. 2017. Evaluating alternative systems of peer review: A large-scale agent-based modelling approach to scientific publication. *Scientometrics*, 113(1):651–671.

John Langford and Mark Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Commun. ACM*, 58(4):12–13.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2281–2287.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Mehdi Moussaid, Juliane E. Kaemmer, Pantelis P. Analytis, and Hansjoerg Neth. 2013. Social Influence and the Collective Dynamics of Opinion Formation. *PLoS ONE*, 8(11):e78433+.

Azzurra Ragone, Katsiaryna Mirylenka, Fabio Casati, and Maurizio Marchese. 2013. On peer review in computer science: Analysis of its effectiveness and suggestions for improvement. *Scientometrics*, 97(2):317–356.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

# Appendicies

## A Consent Message

Before a reviewer or an author enters her reviews or author responses, the following message appears to ask for her consent for data sharing:

> ATTENTION: this time, we plan to do some analytics on anonymized reviews and rebuttal statements, upon the agreement of the reviewers and authors, with the purpose of improving the quality of reviews. The data will be compiled into a unique corpus, which we potentially envisage as a great resource for NLP, e.g. for sentiment analysis and argumentation mining, and made available to the community properly anonymized at earliest in 2 years. We hope to provide data on "how to review" to younger researchers, and improve transparency of the reviewing process in ACL in general.
>
> By default, you agree that your anonymised rebuttal statement can be freely used for research purposes and published under an appropriate open-source license within at earliest 2 years from the acceptance deadline.
>
> Place an 'x' mark in the NO box if you would like to opt out of the data collection.
> [x]: YES
> [ ]: NO

## Analyses on Submissions

We rank n-grams in both accepted and rejected papers according to the *log-likelihood ratio* (LLR) statistic, taking both accepted and rejected papers as one big corpus, respectively. The goal is to find n-grams that occur unusually frequently in one of the two groups, relative to the respective other.

Table 11 shows a few hand-selected n-grams with highest LLR for *accepted papers*; high-LLR n-grams for rejected papers are not presented due to licensing. To filter out noise, we only include n-grams that occur in at least 7 different papers. We can observe some interesting patterns: accepted papers appear to cite recent work, which reflects potential novelty and appropriate comparison to state of the art; tend to use more mathematics (of a particular kind); have an appendix; do significance testing; release code upon publication; and have multiple figures including subfigures.

| Hot n-grams | Possible Interpretation |
|---|---|
| ( 2017 ) | Cite recent work |
| ( $z\|x$ ) | Math |
| artetxe et al | Authors working on a hot topic |
| dozat and manning | Authors of an influential method |
| in the supplementary | Paper has appendix |
| contextualized word representations | Trendy method |
| upon publication . | Code/data will be released |
| statistical significance of | Mathematically rigorous |
| figure 3 ( | Multiple figures with subfigures |

Table 11: Selected 3-grams that distinguish accepted from rejected papers based on the LLR statistics.

## B Statistics on Reviewer Information

In this section, we present some statistics of all 1440 reviewers of ACL-18.

**Country.** The reviewers work in 53 different countries. The top 10 countries where the reviewers work are presented in Fig. 5. The distribution of the reviewer working places is heavily long-tailed: the United States alone contributes 36.9% of all reviewers, followed by China (8.7%), the United Kingdom (7.8%) and Germany (7.6%). Seven countries have more than 50 reviewers, and 19 countries have more than 10 reviewers.

**Affiliation.** The reviewers are from around 700 organisations. But as reviewers use different names to refer to the same organisation (e.g., both MIT and Massachusetts Institute of Technology are used), the real number of organisations can be much lower. The top 10 organisations and their reviewers numbers are presented in Fig. 6. Nine organisations contribute more than 20 reviewers, and 19 organisations contribute more than 10 reviewers.

**Seniority.** Most reviewers (69.9%) do not report their seniority levels. Among those that

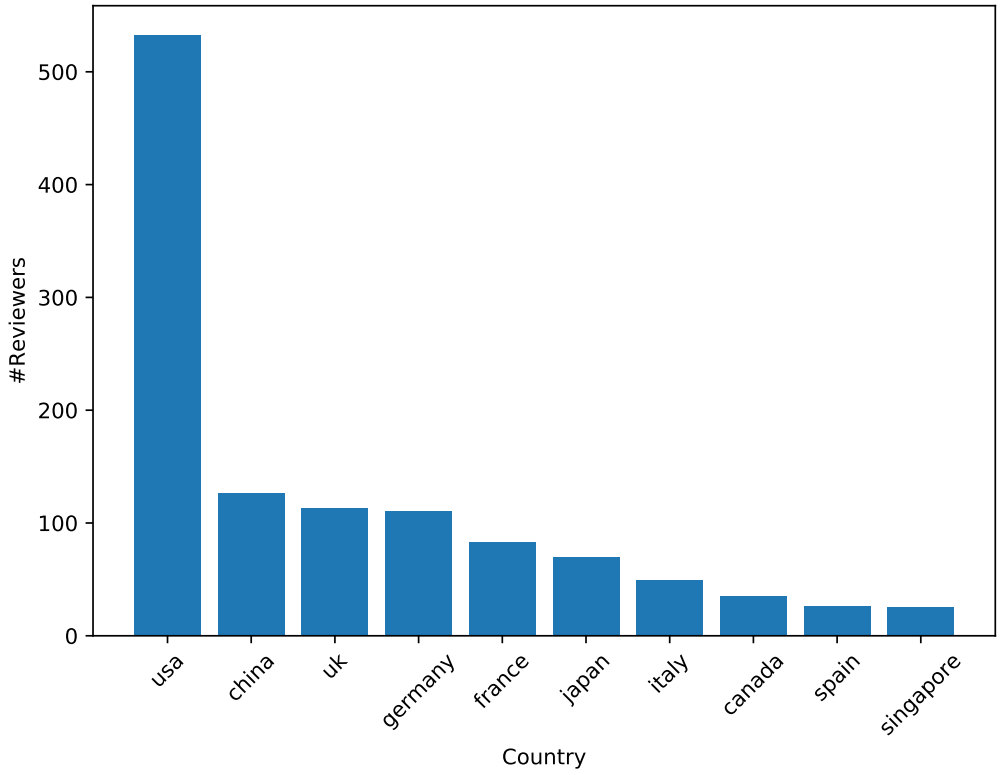


Figure 5: Distribution of countries where reviewers work.

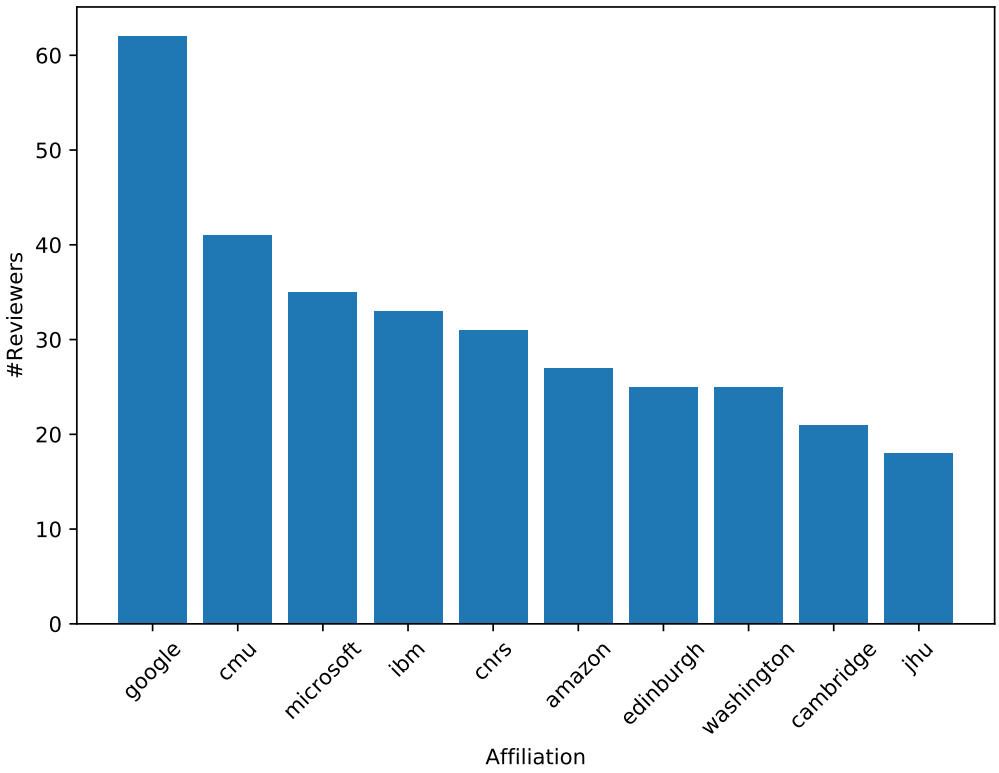


Figure 6: Distribution of organisations where reviewers work.

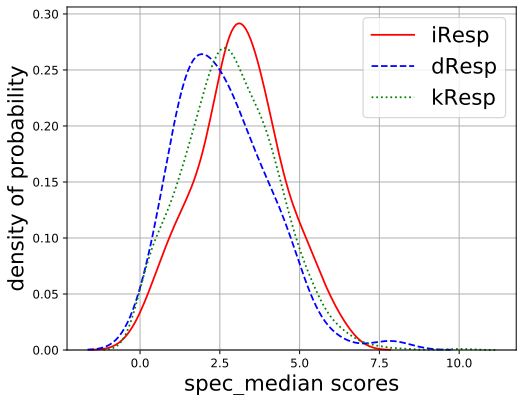


Figure 7: Smoothed distribution of specificity scores.

have reported their seniority, 50.2% are Professors, 27.6% are PhD students, and 22.2% are Post-Doc/Assistant-Professor.

**Gender.** We estimate the gender of the reviewers from their first names, using the tool available at `https://github.com/kensk8er/chicksexer`. 73.4% reviewers are estimated to be male and the rest 26.6% are estimated to be female.

## C Full Results

The precision, recall and F1-scores for each label in both Full and BRD are presented in Table 12 and 13, respectively.

## D Features

The full list of our hand-crafted features is presented in Table 14.

## E Specificity Scores

We tokenize author responses with `nltk`, remove sentences with fewer than 10 tokens and rank the remaining sentences by their specificity scores. All scores are normalized to $[0, 10]$, with higher scores meaning higher specificity. The distribution of the specificity scores for author responses leading to increased, decreased and unchanged scores is illustrated in Fig. 7.

**Top 10 sentences** are presented below[5], and they all receive a specificity score 10.

- *We have already checked it. We can change the sentence in the last paragraph of Section ### to ''Since the proposed method only substituted ### based on ###, then the naturalness of ### using the proposed method is better than ###. This method was used because we have to maintain the context; The result can be more than 100% because we assume that the ### of original was 100% while based on human judgement, there are possibility that the ### of resulting sentences using the proposed method is better than the original one.*
- *###| ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# |*
- *###| ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# |*
- *There are two reasons why we mention that: (i) many papers exist, however, many previous papers made the same (or similar) conclusions, so some are picked up as representatives and (ii) because ### is a high-level conference, it's thought that there was no need to explain too much, and also because there are limited pages, space was wanted to be left to explain the analysis as detailed as possible and put focus on the analysis.*
- *Other external knowledge sources apart from ### do not add much: In principle, all resources we used originate in ###, the difference is the degree of knowledge we use. The novelty in this work does not lie in the use of ### as a knowledge resource but more generally in the principled ### of the classes.*
- *We will include this discussion in the paper. Other ### models (e.g., ###; ###) can in theory predict ###, however, they are not directly applicable to ### since they cannot handle ### representations, i.e., variables can refer to a ### representation (e.g., variable ### refers to an entire proposition and variable ### refers to a segment of meaning).*
- *As noted in our response to reviewer 3 - our results on the ### dataset of ### are on par with the ### model stated in the ### paper provided by reviewer 3 (which is a SOTA non-neural ### model) - although we used a very basic set of features and apply very limited task-specific tuning to our models.*

[5]The examples are anonymized by replacing citations, venues, method names, exact scores, etc. with placeholders; we also include cases where our system has erroneously rated non-text data (i.e. tables).

| Feature Set | INC-p | INC-r | INC-f1 | DEC-p | DEC-r | DEC-f1 | KEEP-p | KEEP-r | KEEP-f1 |
|---|---|---|---|---|---|---|---|---|---|
| spec | .110 | .023 | .035 | 0 | 0 | 0 | .820 | .976 | .892 |
| plt | .043 | .063 | .047 | .029 | .029 | .029 | .824 | .912 | .864 |
| cvc | .020 | .014 | .107 | 0 | 0 | 0 | .824 | .977 | .893 |
| log_leng | .187 | .167 | .154 | 0 | 0 | 0 | .827 | .930 | .874 |
| sim | .013 | .011 | .013 | 0 | 0 | 0 | .810 | .990 | .897 |
| Score | **.331** | .485 | **.386** | **.380** | .527 | .409 | .878 | .790 | .829 |
| All but Score | .142 | .202 | .162 | .025 | .033 | .029 | .820 | .820 | .818 |
| All | .299 | **.555** | .374 | .364 | **.569** | **.438** | **.889** | .757 | .817 |
| Majority Baseline | 0 | 0 | 0 | 0 | 0 | 0 | .823 | **1** | **.903** |
| Random Baseline | .100 | .332 | .154 | .076 | .334 | .123 | .825 | .332 | .474 |

Table 12: Macro F-1 scores on Full. All results are averaged over 5000 repeats of 10-fold cross validation.

| Feature Set | INC-p | INC-r | INC-f1 | DEC-p | DEC-r | DEC-f1 | KEEP-p | KEEP-r | KEEP-f1 |
|---|---|---|---|---|---|---|---|---|---|
| spec | .119 | .101 | .102 | 0 | 0 | 0 | .804 | .956 | .872 |
| plt | .100 | .012 | .022 | .020 | .014 | .017 | .804 | .982 | .883 |
| cvc | .033 | .020 | .025 | 0 | 0 | 0 | .803 | .988 | .885 |
| log_leng | .180 | .229 | .184 | 0 | 0 | 0 | .811 | .879 | .840 |
| sim | .096 | .133 | .110 | 0 | 0 | 0 | .805 | .927 | .861 |
| Score | **.313** | .556 | **.394** | .377 | .356 | .302 | .851 | .743 | .792 |
| All but Score | .205 | .331 | .231 | .050 | .011 | .018 | .801 | .768 | .780 |
| All | .295 | **.570** | .376 | **.387** | **.548** | **.418** | **.875** | .710 | .782 |
| Majority Baseline | 0 | 0 | 0 | 0 | 0 | 0 | .802 | **1** | **.890** |
| Random Baseline | .117 | .333 | .173 | .082 | .335 | .131 | .802 | .333 | .470 |

Table 13: Macro F-1 scores on BRD. All results are averaged over 5000 repeats of 10-fold cross validation.

| Feature set | Features |
|---|---|
| Score | self_before, self_conf, oth_max, oth_min, oth_mean, oth_median, oth_std, oth_conf_max, oth_conf_min, oth_conf_mean, oth_conf_median, oth_conf_std, oth_mean-self, oth_median-self, oth_max-self, self-oth_min, oth_conf_std, all_max, all_min, all_mean, all_median, all_std, self_before**2, all_mean-self, all_max-self, all_median-self, self-all_min |
| spec | spec_max, spec_min, spec_mean, spec_median, spec_std |
| cvc | cvc_max, cvc_min, cvc_mean, cvc_median, cvc_std |
| plt | plt_max, plt_min, plt_mean, plt_median, plt_std |
| log_leng | Logarithm of the token number of the author response |
| sim | Cosine similarity of the embeddings of a review and its corresponding author response |

Table 14: The full list of hand-crafted features.

- *Although the models used are general to all seq2seq generation problems, the heuristics we used to select ### are specific to generating the ### (take for example, the heuristic based on ### - it was motivated by the fact that ### have a higher readability, hence the network has to focus towards better readable information in the ### in order to generate ###).*

- *Because the size of the training data for ### task is very small, ### instances for ### task and ### instances for ### task, whereas the number of the parameters of the whole network is very big, we pre-training the ### network based on ###, released for ### task, and pre-training the ### network based on the training data for ### task.*

- *Related workshop and share tasks, including ### (collocated with ###), ### (collocated with ###), ### (collocated with ###), and ### (collocated with ###), show a great potential on applying NLP technologies to the ### domain.*

**Bottom 10 sentences** are presented below. Their specificity scores are all smaller than 0.001.

- *It would be a little difficult to build this connection.*

- *It is not accurate and we will use 'obvious' instead.*

- *We are not quite sure which part is not identical.*

- *I will check that again and will write it as you said*

- *Therefore we can see that they have no relation with each other.*

- *We will try to do this in our future work.*

- *But we do not see this as a weakness of our approach.*

- *That is why we do not do that in the first submission.*

- *So this is really true for all the "models".*

- *Thank you very much for the reviews and for the very useful*

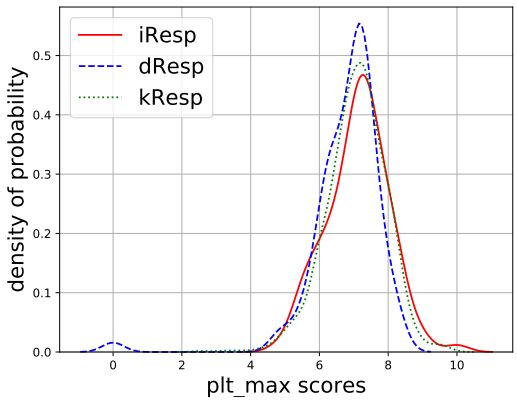


Figure 8: Smoothed distribution of politeness scores.

## F Politeness Scores

We use the politeness scorer to rate the same set of sentences as in the specificity evaluation. We normalize all politeness scores to [0,10], with higher values meaning higher politeness. The distribution of the politeness scores is illustrated in Fig. 8. **Top 10 sentences** and their politeness scores are presented below.

- *(9.6) We thank this reviewer for his helpful comments that help improving the paper.*

- *(9.5) Thanks for the suggestion, we found that in many cases the two sentences that are separated by ### also have similar patterns to ###, and the size of the dataset would be too small to train a representative ### model if we only picked out the separate sentences examples.*

- *(9.5) Thank you for the helpful suggestion of including more qualitative results to more thoroughly understand the proposed approach.*

- *(9.4) We again thank the reviewer for the detailed and carefully constructed review and assure that the main concerns raised by the reviewer are fixable and we will fix them in the final version of the paper.*

- *(9.4) Meanwhile, thanks for your suggestion for more in-depth discussion on ###*

- *(9.3) We apologize for this error, and will correct this in the final version of the paper upon acceptance.*

- *(9.3) An interesting alternative approach would be the one proposed by the reviewer, but we chose this model because we wanted to encourage the model to aggregate information from a variety of positions, and in our experience ### has trouble learning to ### in this way because by design ### focuses on one position only.*

- *###| ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# |*

- *###| ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# | ##.# |*

- *(9.2) Depends on the task and the characteristic of two datasets, each proposed method shows its effectiveness, e.g., the ### using the ### between two entities is appropriate for the ### task since ### is systematically organized.*

**Bottom 10** sentences and their politeness scores are presented below.

- *(1.7) By comparing ### with ###-, we know whether employing a ### helps; By comparing ### with ###, we know whether employ a ### helps; By comparing ### with ###, we know whether the ### helps.*

- *(2.2) $\#\#\# = \#\#\# * \#\#\#$, where ### is a matrix of n samples with ### features followed by ### features, hence the size of $\#\#\#$ is $\#\#\#$.*

- *(2.3) In other words, our coverage is ### times larger than theirs, so our proposed system can deal much better with the noise when learning ###.*

- *(2.4) And another difference lies in the ### layer, which contains ###, so when we process ### in ### independently which encourages our model to learn diverse features.*

- *(2.4) We will implement their method on our corpora and make some comparison with our method in the next version of our manuscript.*

- *(2.4) We are not giving up ### nor are we claimining that ### is more powerful.*

- *(2.5) If our paper is accepted we will make sure additional relevant technical details are added.*

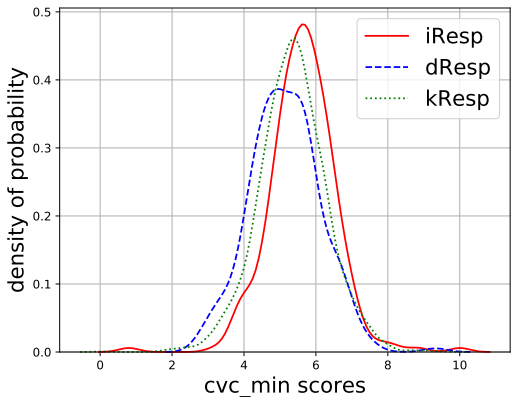


Figure 9: Smoothed distribution of the convincingess scores.

- *(2.6) In response to your general remark: we can see how our discussion and conclusions would lead a reader to conclude that; rather, this paper is an exploration in an area that is, as you say, worth exploring.*

- *(2.7) Our main contribution is introduction of ### without requiring neither supervision nor feature engineering.*

- *(2.7) The most salient problem encountered in our system is that a user might change ###, also brought up by R3 (Please refer to our response to weakness4 of R3).*

## G Convincingness Scores

We use the convincingness scorer to rate the same set of sentences as in the previous two studies. The convincingness scores are normalized to $[0, 10]$, with higher values meaning higher convincingness. The distribution of the convincingness scores is illustrated in Fig. 9. **Top 10 sentences** in terms of convincingness are presented below. All top 10 sentences' convincingness scores are above 9.8.

- *In the revision, we perform the evaluation of the model with ### and ###, respectively.*

- *Deepening the ### system would inevitably increase model parameters, and slow the training and decoding, which is not what we expect.*

- *A technical document is defined as the document that assumes sufficient background*

*knowledge and familiarity with the key technical or central/important terms in the document.*

- *As reported in our paper, the success rate of our optimization algorithm is ### while, on average, only ###% of words are altered.*
- *The focus of this work is not a comparison of ### methods with ### methods, but how to mitigate the lack of labeled data problem in learning of a ### model.*
- *Our model works well on datasets that are deemed small for deep architectures to work and belong to special domains for which ### is not possible.*
- *We conduct t-test and get the p value as ###, which shows good agreement.*
- *Particularly, we will strive to improve the presentation quality and to make the draft more readable and better organized for more potential readers.*
- *Furthermore, ### can help ### to alleviate the performance degradation by ###.*
- *The ### experiments in Section ### show that our ### framework can achieve higher accuracy than the methods that rely on the same set of resources, while the state-of-the-art ### methods also require some other resources.*

**Bottom 10 sentences** in terms of convincingness scores are presented below[6]. Their convincingness scores are all below 0.01.

- *"Weakness 3:""why ... report on ... the '###' if you then dismiss it""""*
- *It is **not** used in the **testing** (###).*
- *Annotator 1: "Are you a citizen?" No =¿ Answer: No*
- *"Rev: ""It seems that ...""""*
- *"Weakness 3:""how did you learn the embeddings? ... ### model? How"*
- *Please refer to the reply regarding Weakness argument 1 in Review 1.*
- *"Are you over 21?" Yes =¿ Answer: Yes*
- *Please see our reply to Review 1's weakness argument 3.*
- [*Please see our response to R2's argument 3*]
- *We are sorry we didn't explain the notation.*

[6]Note the large number of references to other responses and to the original reviews