# A Reproduction Study of the Human Evaluation of Role-Oriented Dialogue Summarization Models

**Mingqi Gao, Jie Ruan, Xiaojun Wan**
Wangxuan Institute of Computer Technology, Peking University
{gaomingqi, wanxiaojun}@pku.edu.cn
ruanjie@stu.pku.edu.cn

## Abstract

This paper reports a reproduction study of the human evaluation of role-oriented dialogue summarization models, as part of the ReproNLP Shared Task 2023 on Reproducibility of Evaluations in NLP. We outline the disparities between the original study's experimental design and our reproduction study, along with the outcomes obtained. The inter-annotator agreement within the reproduction study is observed to be lower, measuring 0.40 as compared to the original study's 0.48. Among the six conclusions drawn in the original study, four are validated in our reproduction study. We confirm the effectiveness of the proposed approach on the overall metric, albeit with slightly poorer relative performance compared to the original study. Furthermore, we raise an open-ended inquiry: how can subjective practices in the original study be identified and addressed when conducting reproduction studies?

## 1 Introduction

Reproducibility has gained significant attention within the field of Natural Language Processing (NLP) in recent years. This paper presents a reproduction study focused on the human evaluation of role-oriented dialogue summarization models. The study was conducted as part of the ReproNLP Shared Task 2023, which aims to foster reproducibility in NLP evaluations. Our participation in Track C involved conducting a reproduction study specifically targeting the human evaluation component described in the work by Lin et al. (2022), which is one of the five papers included in this track. The shared dataset used in this track originates from the ReproHum project [1], which employs a multi-lab paradigm to assess reproducibility in NLP.

[1] https://reprohum.github.io/

Role-oriented dialogue summarization aims to generate summaries tailored to various roles within a conversation. For instance, in the context of a customer service chat, distinct summaries can be generated for the user's and the agent's utterances. The original research paper introduced an approach that leverages role interaction to effectively integrate the content of other roles into the summary pertaining to a specific role (Lin et al., 2022). The aforementioned study empirically demonstrated the effectiveness of the proposed approach in comparison to baseline methods through both automatic and human evaluations. In this study, we specifically concentrate on the human evaluation aspect.

## 2 Experimental Design

### 2.1 Original experiment

Lin et al. (2022) applied the proposed approach to two popular sequence-to-sequence models: PGN (See et al., 2017) and BERTAbs (Liu and Lapata, 2019). The baseline dialogue summarization models are denoted as **PGN-multi** and **BERT-multi**, and the models with the role interaction approach are noted as **PGN-both** and **BERT-both**. The human evaluation was conducted on CSDS (Lin et al., 2021), a Chinese customer service dialogue summarization dataset.

**Selection of evaluation samples**. From the test set of CSDS, 100 dialogues were randomly chosen as evaluation samples. Each dialogue is associated with two reference summaries, one for the user and one for the agent. A model also generated summaries for both the user and the agent. For each reference summary, four model-generated summaries (PGN-multi, BERT-multi, PGN-both, and BERT-both) were evaluated by human annotators. Notably, the source dialogues were excluded from the human evaluation process.

**Participating annotators and compensation**.



*Proceedings of the Third Workshop on Human Evaluation of NLP Systems*,
pages 124–129, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-088-5_010

Three Chinese graduate students, all proficient in Chinese, volunteered as annotators for this evaluation. These participants were not remunerated for their involvement.

**Evaluation dimensions and criteria**. Given a reference summary, a model-generated summary was evaluated on three dimensions: *Informativeness*, *non-redundancy*, and *fluency*. Specifically, the annotators were asked to rate each sentence in the summary on a Likert scale from 0 to 2.

*Informativeness*: The reference summary is composed of multiple sentences, and the annotators were asked to determine whether the information of each sentence in the reference summary is extracted by the model-generated summary. For each sentence in the reference summary, the rule is as follows:

- 0 if most of its content is not extracted by the model-generated summary.
- 1 if some of its content is extracted.
- 2 if basically all of its content is extracted.

*Non-redundancy*: The model-generated summary is also composed of multiple sentences, and the annotators were asked to determine whether the information of each sentence in the model-generated summary is redundant. For each sentence in the model-generated summary, the rule is as follows:

- 0 if its content is not in the reference summary.
- 1 if its content is in the reference summary but there is redundancy compared to the reference summary.
- 2 if the content is basically the same.

*Fluency*: For each sentence in the model-generated summary, the rule is as follows:

- 0 if it has more grammatical errors or misspellings, or if the statement is incomprehensible.
- 1 if it has minor grammatical errors or typos, or if the expression is more colloquial.
- 2 if the expression is fluent, free of grammatical errors and misspellings, and semantically completed.

**Annotation interface**. The reference summaries and model-generated summaries were presented to annotators using an Excel sheet, and they filled in the ratings in the specified places as shown in Figure 1. To ensure impartiality, the names of the summarization models were withheld from the annotators, and the order of the model-generated summaries was randomized.

**Annotation procedure**. Annotators were asked to read the evaluation instructions before annotation. Initially, all three annotators independently annotated the first 10 samples (ID 0-9). After a moderate level of inter-annotator agreement was attained, they were allowed to continue annotation. The remaining 90 samples were divided equally into thirds. The remaining 90 samples were evenly divided into thirds. Annotator #1 was assigned samples with ID 10-39, annotator #2 received samples with ID 40-69, and annotator #3 handled samples with ID 70-99.

**Inter-annotator agreement**. The results of the first 10 samples were used to compute inter-annotator agreement. All per-sentence scores given by an annotator on all three dimensions are flattened into a list. The Cohen's kappa (Cohen, 1960) was computed between every two annotators with the script in the scikit-learn library[2], and the average of the kappa scores was considered as the final inter-annotator agreement.

**Post-processing, calculation, and significance testing**. To normalize the scores to a range of 0 to 1, they were divided by 2. For the first 10 samples, the annotations of the annotator with the most expertise were selected as the final results. For each of the three dimensions, the per-sentence scores of the summary were averaged as the score of the summary. In addition, the average of the summary-level scores of the three dimensions was calculated as an "Overall" score for a summary. The model's score was obtained by averaging the scores of its generated summaries. A paired t-test was conducted to assess the significance between the scores of summaries generated by two models.

### 2.2 Reproduction experiment

The reproduction experiment utilized the same Excel sheet for annotation as the original study, which encompassed identical samples for evaluation. Furthermore, the evaluation instructions were also pro-

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

| Dialogue ID | Reference Summary | The summary to be evaluated | Informativeness | Non-redundancy | Fluency |
|---|---|---|---|---|---|
| 0 | Users ask about methods other than cell phone verification to change their passwords. Users ask how they can log in with a password without cell phone number verification. Users ask if they can buy something without verification. | The user said that the previous cell phone number was canceled, the password was forgotten, and the cell phone verification code was needed to change the password. | | | |
| | | The user said that the previous cell phone number was canceled, the password was forgotten, and the cell phone verification code was needed to change the password. | | | |
| | | The user wants to change the password. The user asks if a verification code is required. | 1,0,0 | 2,0 | 2,1 |
| | | The user says he has forgotten his previous cell phone number and wants to change his password. The user asks if he can buy something. The user says he can't change his password by email. | | | |
| | Customer service allows the user to provide the number and give it to the commissioner to call back to solve the problem. Customer service helps feedback the user's problem to the commissioner. Customer service asks the user to wait patiently. | Customer service helps the user upgrade the commissioner will be tomorrow [number] o'clock before the user calls back to facilitate the provision of the user to bind the cell phone number. Customer service answer has been feedback commissioner, please keep the user's phone open. | | | |
| | | Customer service helps user upgrade specialist will call back before tomorrow [number] o'clock to facilitate the provision of the user to bind the cell phone number. | | | |
| | | Customer service replied to help the user to upgrade the commissioner to deal with, and told the user that the commissioner will call the user back by tomorrow [number] o'clock. Customer service answers the user can see the previous cell phone number. | | | |
| | | Customer service answers to help the user feedback commissioner and call the user back by tomorrow [number] o'clock. | | | |

Figure 1: Annotation interface. The text actually presented to annotators is in Chinese, and the translated version is shown here.

vided. With these materials, we were able to set up most of the experiment in the exact same way as the original. Nonetheless, certain variations were introduced in the reproduction experiment, which is outlined below. For more detailed information, please see the Human Evaluation Sheet (HEDS) file in supplementary materials.

**Participating annotators and compensation**. Recruiting unpaid volunteers as annotators proved to be challenging. Following discussions with the organizers of the ReproHum project, we recruited three participants who met the same requirements as those in the original experiment and provided them with compensation of 12.24 EUR per hour.

**Annotation procedure**. We cannot know what the original experiment would have done if the three annotators had not reached a moderate level of agreement on the first 10 samples because this did not actually happen. In consultation with the organizers of the ReproHum project, we determined that all annotators would continue with the annotation process, regardless of whether a moderate agreement was reached on the first 10 samples or not.

**Post-processing, calculation, and significance testing**. It is subjective to determine which participant was most knowledgeable on this task. Given the challenging nature of reproduction, the organizers of the ReproHum project asked us not to copy the original practices to post-process the first 10 samples. they proposed that we calculate separate results using each of the following five methods (referred to as different reproduction settings):

- Repr1: With only annotator #1 representing each sentence in the first 10 samples (as if #1 had been selected).
- Repr2: With only annotator #2 representing each sentence in the first 10 samples (as if #2 had been selected).
- Repr3: With only annotator #3 representing each sentence in the first 10 samples (as if #3 had been selected).
- Repr4: With the mean of annotator responses representing each sentence in the first 10 samples (i.e., [0,1,2] $\Rightarrow$ 1.00, [0,0,2] $\Rightarrow$ 0.67).
- Repr5: With the median of annotator responses representing each sentence in the first 10 samples (i.e., [0,1,2] $\Rightarrow$ 1, [0,0,2] $\Rightarrow$ 0).

In addition, our reproduction experiments began after the protocol was approved by the ethics committee.

## 3 Results and Discussion

### 3.1 Inter-annotator agreement (IAA)

The initial study reported an IAA of 0.48, while our reproduction experiment yielded a slightly lower IAA of 0.40. The IAA observed in the original study can be classified as moderate (0.41-0.60), and

| Original | | | | | Repr3 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Info | Non-Red | Flu | Overall | | Info | Non-Red | Flu | Overall |
| PGN-multi | **0.69**/0.65 | 0.54/0.55 | 0.70/0.79 | 0.64/0.66 | PGN-multi | 0.68/0.62 | 0.60/0.56 | 0.82/**0.90** | 0.70/0.69 |
| PGN-both | 0.66/**0.69** | **0.58/0.59*** | **0.73/0.81** | **0.66/0.70*** | PGN-both | 0.68/**0.66*** | **0.61/0.59** | **0.84**/0.89 | **0.71/0.71** |
| BERT-multi | 0.58/0.56 | **0.66/0.61** | 0.84/0.87 | 0.69/0.68 | BERT-multi | 0.57/0.52 | **0.67**/0.56 | **0.91**/0.89 | **0.71**/0.66 |
| BERT-both | **0.62*/0.60*** | 0.62/0.60 | **0.85**/0.87 | **0.70/0.69** | BERT-both | **0.59/0.56** | 0.62/**0.58** | 0.87/0.89 | 0.69/**0.68** |
| Repr1 | | | | | Repr4 | | | | |
| | Info | Non-Red | Flu | Overall | | Info | Non-Red | Flu | Overall |
| PGN-multi | 0.68/0.62 | 0.60/0.55 | 0.80/0.89 | 0.69/0.69 | PGN-multi | 0.68/0.62 | 0.59/0.55 | 0.81/0.89 | 0.69/0.69 |
| PGN-both | **0.69/0.68*** | **0.61/0.60*** | **0.83**/0.89 | **0.71/0.72*** | PGN-both | 0.68/**0.67*** | **0.60/0.59** | **0.83**/0.89 | **0.71/0.72*** |
| BERT-multi | 0.57/0.51 | **0.67**/0.57 | **0.90**/0.88 | **0.71**/0.66 | BERT-multi | 0.56/0.51 | **0.67**/0.56 | **0.90**/0.88 | **0.71**/0.65 |
| BERT-both | **0.60/0.56*** | 0.63/**0.58** | 0.86/0.88 | 0.70/**0.68** | BERT-both | **0.59/0.56*** | 0.62/**0.58** | 0.87/**0.89** | 0.69/**0.67** |
| Repr2 | | | | | Repr5 | | | | |
| | Info | Non-Red | Flu | Overall | | Info | Non-Red | Flu | Overall |
| PGN-multi | 0.67/0.62 | 0.58/0.55 | 0.80/**0.90** | 0.68/0.69 | PGN-multi | 0.68/0.62 | 0.59/0.55 | 0.81/**0.90** | 0.69/0.69 |
| PGN-both | 0.67/**0.66*** | **0.60/0.58** | **0.83**/0.89 | **0.70/0.71** | PGN-both | 0.68/**0.67*** | **0.61/0.59*** | **0.83**/0.89 | **0.70/0.72** |
| BERT-multi | 0.56/0.51 | **0.67**/0.56 | **0.91**/0.89 | **0.71**/0.65 | BERT-multi | 0.57/0.51 | **0.67**/0.56 | **0.90**/0.88 | **0.71**/0.65 |
| BERT-both | **0.58/0.55** | 0.61/**0.57** | 0.87/0.89 | 0.69/**0.67** | BERT-both | **0.59/0.56** | 0.62/**0.58** | 0.87/**0.89** | 0.69/ **0.67** |

Table 1: Human evaluation results in the original experiment and the reproduction experiment. Two values separated by a slash in a cell are scores for user-oriented summary and agent-oriented summary. * denotes that the enhancement achieved by utilizing role interactions, compared to the *multi* baseline, is statistically significant ($p < 0.05$). The original results are taken from Lin et al. (2022). "Repr#" is defined in Section 2.2.

| | Original | Reproduction | Confirmation |
|---|---|---|---|
| 1 | For PGN models, applying role interactions could reduce redundancy. | For PGN models, applying role interactions could reduce redundancy. | Confirmed. |
| 2 | For PGN models, applying role interactions could maintain a comparable performance of informativeness. | For PGN model, applying role interactions could partially improve informativeness. | Confirmed. The relative performance of the proposed approach in the experiment is slightly better than the original. |
| 3 | For PGN models, applying role interactions could improve fluency. | For PGN models, applying role interaction could maintain a comparable performance of fluency. | Not confirmed. The relative performance of the proposed approach in the experiment is worse than the original. |
| 4 | For BERTAbs models, applying role interactions could improve informativeness. | For BERTAbs models, applying role interactions could improve informativeness. | Confirmed. |
| 5 | For BERTAbs models, applying role interactions could add redundancy. | For BERTAbs models, applying role interactions could maintain a comparable performance of non-redundancy. | Not confirmed. The relative performance of the proposed approach in the experiment is better than the original. |
| 6 | Applying role interactions is effective in terms of the overall metric. | For PGN models, applying role interactions is effective in terms of the overall metric. | Confirmed. The relative performance of the proposed approach in the experiment is slightly worse than the original. |

Table 2: The conclusions in the original experiment and the reproduction experiment. The *Confirmation* column shows whether the conclusion is confirmed in the reproduction experiment or not and how the relative performance changed in the reproduction experiment. **Note** that relative performance refers to the results of the proposed approach relative to the baseline model.

| All reproduction settings (Repr1, 2, 3, 4, 5) | | | | | Original vs. Repr3 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Info | Non-Red | Flu | Overall | | Info | Non-Red | Flu | Overall |
| PGN-multi | 0.74/0.00 | 1.58/0.90 | 1.16/0.68 | 1.14/0.00 | PGN-multi | 1.46/4.71 | 10.49/1.80 | 15.74/12.98 | 8.93/4.43 |
| PGN-both | 1.16/1.40 | 1.01/1.34 | 0.60/0.00 | 0.87/0.85 | PGN-both | 2.98/4.43 | 5.03/0.00 | 13.97/9.38 | 7.28/1.41 |
| BERT-multi | 1.08/0.98 | 0.00/0.89 | 0.68/0.69 | 0.00/0.94 | BERT-multi | 1.73/7.39 | 1.50/8.52 | 7.98/2.27 | 2.85/2.98 |
| BERT-both | 1.34/0.90 | 1.27/0.86 | 0.58/0.56 | 0.72/0.91 | BERT-both | 4.94/6.88 | 0.00/3.38 | 2.32/2.27 | 1.43/1.46 |
| Original vs. Repr1 | | | | | Original vs. Repr4 | | | | |
| | Info | Non-Red | Flu | Overall | | Info | Non-Red | Flu | Overall |
| PGN-multi | 1.46/4.71 | 10.49/0.00 | 13.29/11.87 | 7.50/4.43 | PGN-multi | 1.46/4.71 | 8.82/0.00 | 14.53/11.87 | 7.50/4.43 |
| PGN-both | 4.43/1.46 | 5.03/1.68 | 12.78/9.38 | 7.28/2.81 | PGN-both | 2.98/2.93 | 3.38/0.00 | 12.78/9.38 | 7.28/2.81 |
| BERT-multi | 1.73/9.32 | 1.50/6.76 | 6.88/1.14 | 2.85/2.98 | BERT-multi | 3.50/9.32 | 1.50/8.52 | 6.88/1.14 | 2.85/4.50 |
| BERT-both | 3.27/6.88 | 1.60/3.38 | 1.17/1.14 | 0.00/1.46 | BERT-both | 4.94/6.88 | 0.00/3.38 | 2.32/2.27 | 1.43/2.93 |
| Original vs. Repr2 | | | | | Original vs. Repr5 | | | | |
| | Info | Non-Red | Flu | Overall | | Info | Non-Red | Flu | Overall |
| PGN-multi | 2.93/4.71 | 7.12/0.00 | 13.29/12.98 | 6.04/4.43 | PGN-multi | 1.46/4.71 | 8.82/0.00 | 14.53/12.98 | 7.50/4.43 |
| PGN-both | 1.50/4.43 | 3.38/1.70 | 12.78/9.38 | 5.86/1.41 | PGN-both | 2.98/2.93 | 5.03/0.00 | 12.78/9.38 | 5.86/2.81 |
| BERT-multi | 3.50/9.32 | 1.50/8.52 | 7.98/2.27 | 2.85/4.50 | BERT-multi | 1.73/9.32 | 1.50/8.52 | 6.88/1.14 | 2.85/4.50 |
| BERT-both | 6.65/8.67 | 1.62/5.11 | 2.32/2.27 | 1.43/2.93 | BERT-both | 4.94/6.88 | 0.00/3.38 | 2.32/2.27 | 1.43/2.93 |

Table 3: CV*s among all reproduction settings (Repr1, 2, 3, 4, 5) and CV*s between scores in the original experiment and scores in the reproduction experiment with a specific setting. Two values separated by a slash in a cell are scores for user-oriented summary and agent-oriented summary.

the slightly lower IAA in the reproduction study falls near the boundary between the moderate and fair levels. There is not much difference between the two. Nevertheless, it might be more reasonable to calculate the IAA independently for each of the three evaluation dimensions, but only an overall IAA was reported in the original study.

### 3.2 Side-by-side comparison of conclusions

Table 1 presents the human evaluation results of various models in both the original experiment and the reproduction experiment conducted under different settings (Repr1, 2, 3, 4, 5). It is evident that the outcomes of the reproduction experiment exhibit minor divergence across the different settings. The original paper posits six conclusions, each of which can be assessed for confirmation based on the results of the reproduction experiment, as depicted in Table 2. Notably, **four out of the six conclusions are substantiated**.

Furthermore, our analysis centers on the variations observed in the relative performance of the proposed approach between the reproduction experiment and the original experiment. In certain aspects, such as the informativeness of the summaries generated by PGN models, the reproduction experiment demonstrates an improvement over the original experiment. Conversely, in other aspects, the relative performance of the proposed approach is inferior to that of the original experiment. In particular, the fifth conclusion from the original experiment, as stated in Table 2, highlights a drawback of the proposed approach. However, this drawback is not supported by the findings of the reproduction experiment. As for the sixth conclusion from the original experiment, **the effectiveness of the proposed approach is confirmed in terms of the overall metric, although the relative performance in the reproduction experiment exhibits a slight decline in comparison to the original experiment.**

### 3.3 Quantifying the difference

To quantify the disparities between the outcomes of the original experiment and the reproduction experiment, as well as the variations in the results across different settings in the reproduction experiment, we employ two statistical measures: the small-sample coefficient of variation (CV*) and Spearman's $\rho$.

A lower value of CV* corresponds to a smaller discrepancy, rendering it a quantifiable metric for assessing the reproducibility of numerical scores (Belz et al., 2022). Table 3 demonstrates that **the CVs between scores obtained in the original experiment and those obtained in the reproduction experiment with a specific setting are considerably larger than the CVs observed among different reproduction settings.** This finding suggests that the variations introduced by distinct reproduction settings, specifically the methods employed for post-processing the initial 10 samples, have a relatively minor impact on the results.

In Table 4, we present the system-level Spearman's rank correlation between the original experiment and the reproduction experiment. The con-

| | Info | Non-Red | Flu | Overall |
|---|---|---|---|---|
| Repr1 | 0.80/1.00 | 1.00/0.20 | 0.80/-0.94 | 0.32/0.40 |
| Repr2 | 0.95/1.00 | 1.00/0.20 | 0.80/-0.82 | 0.40/0.40 |
| Repr3 | 0.95/1.00 | 1.00/-0.11 | 0.80/-0.82 | -0.32/0.40 |
| Repr4 | 0.95/1.00 | 1.00/0.20 | 0.80/-0.54 | 0.00/0.40 |
| Repr5 | 0.95/1.00 | 1.00/0.20 | 0.80/-0.83 | 0.11/0.40 |

Table 4: Spearman's $\rho$ between the scores of four models in the original experiment and the reproduction experiment with a specific setting. Two values separated by a slash in a cell are scores for user-oriented summary and agent-oriented summary.

siderable variation across different dimensions can be attributed to the limited number of comparable systems in this study. Therefore, it is better to use CV* to measure the differences in this case.

## 4 Conclusion

We present a reproduction study focused on evaluating dialogue summarization models through human evaluation. The successful execution of our reproduction experiment was facilitated by the collaboration with ReproNLP organizers and the utilization of materials provided by the original authors. As a result, we have drawn the following conclusions:

- The inter-annotator agreement in our reproduction study was found to be lower, with a value of 0.40 compared to the original study's 0.48.
- Four out of the six conclusions reached in the original study were confirmed through our reproduction study.
- Our findings affirm the effectiveness of the proposed approach in terms of the overall metric; however, the relative performance was slightly inferior in the reproduction study.
- The utilization of different post-processing methods for the first 10 samples yielded minor variations in the final results.

One intriguing query that arises in the context of our reproduction study pertains to the identification and handling of subjective practices that may have been employed in the original study. Specifically, we explore the different post-processing methods of annotation results from the initial 10 samples in this experiment. Despite the limited impact of varying treatments on the ultimate outcome, the underlying concern persists. Notably, if subjective practices are embedded within the core of the original experiment, the potential simulation of multiple possibilities can significantly amplify the scale of the experiment. This matter merits further investigation and remains an avenue for future research.

## References


Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.