Automatic Evaluation of Consistency in Dialogue System Responses

対話システムの応答における一貫性の自動評価

Shiki Sato

Graduate School of Information Sciences
Tohoku University

This dissertation is submitted for the degree of
Doctor of Information Science

January 2024

Acknowledgements

本研究を進めるなかで、本当に多くの方々にご助力いただきました。ここに深謝の意を表します。

主指導教官である鈴木潤教授には、学部 4 年次での研究室配属から約 6 年間にもわたる研究活動において、俯瞰的でありながら親身なご助言をいつもいただき、研究を進めていくなかで躓いた際にはいつも救われました。詳細な技術面でも多数ご助言をいただき、技術者としても成長することができたと考えております。心より感謝申し上げます。

お忙しいなか、本論文の審査委員をお引き受けくださり、適切なご助言を賜りました北村喜文教授ならびに伊藤彰則教授に感謝申し上げます。

乾健太郎教授には、研究室配属当初から、研究内容に対するご助言はもちろん、研究という営みの意味ややりがい、楽しさに気づかせてくださるようなお言葉を多数いただき、博士後期課程進学に至るモチベーションを与えてくださいました。また学生生活に関するものも含む様々なご相談に乗っていただき、不安を抱え込むことなく研究活動に専念することができました。心より感謝の意を表します。

坂口慶祐准教授には、ご多忙のところいつも相談に対して丁寧に対応していただき、特に論文執筆等で思い悩んでいる際にたくさんのアドバイスをいただきました。おかげをもちまして、本論文を執筆することができました。御礼申し上げます。

赤間怜奈助教には、修士課程からの約 5 年という長い期間、日々の研究活動から論文執筆、学生生活に至るまで、本当に色々な場面で温かくかつ辛抱強く支えていただきました。自然言語処理のなかでも様々な要素を含む総合格闘技とも言われるこの研究分野で本研究を成し遂げることができたのは、対話システム研究に携わる先輩である赤間助教のご助力あってこそと考えており、感謝の気持ちで一杯です。

奈良先端科学技術大学院大学の大内啓樹助教には、私の拙い言葉にいつも真摯に耳を傾けていただき、対話研究に興味を持つきっかけをいただきました。その後も、お忙しいなかでも常に気にかけていただき本研究を支えていただきました。感謝の念に堪えません。

徳久良子准教授には、本研究に関するアドバイスのほかにも、いつも思いやりのこもったお言葉を折に触れてかけていただき、心も健やかに研究に邁進することができました。誠にありがとうございました。

学会やシンポジウムで議論してくださった皆様のお言葉一つ一つが、研究を発展させていくうえで非常に参考になりました。議論させていただいたすべての皆様に拝謝します。

Tohoku NLP の皆様には、日々の研究に対するご助言をいただくとともに、様々な交流をさせていただくことで、孤独を感じない研究生活を送ることができました。心よりお礼申し上げます。

最後に、当初は考えていなかった博士後期課程進学という決断を尊重し、ときに励ましの言葉をかけてくれた家族に深く感謝します。

Abstract

Avoiding the generation of responses that contradict the preceding context is a significant challenge for open-domain dialogue response generation systems (RGSs). Automatic evaluation for the ability to avoid contradictory responses (Consistency-Awareness, CA) of RGSs has the potential for the improvement of CA because of its high reproducibility and low cost. The development of an automatic evaluation framework that is reliable, i.e., highly correlated with human CA evaluations, is critical to effectively improving the CA of RGSs.

In this study, we first identify which of the existing CA automatic evaluation approaches are effective for the efficient improvement of RGSs' CA. Our experiments confirmed that automatic CA evaluation based on the probability-based approach may have a limitation in its evaluation effectiveness; we need to employ generation-based CA evaluation methods.

Based on this result, we address the improvement of the effectiveness of generation-based automatic evaluation by enhancing the performance of contradiction detectors by augmenting their training data with real RGS-generated contradictory responses. Furthermore, in order to evaluate more practical CA of RGSs, we propose a framework for automatic CA evaluation of n-best candidates assuming the removal of contradictory response candidates of RGS by post-processing.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Open-domain dialogue response generation systems (RGSs) have attracted attention in various fields, including medicine and education, and are the subject of active research and development. In particular, deep neural network-based RGSs, which have been studied rapidly in recent years against the backdrop of the development of deep learning technology, are known to be able to generate fluent responses to dialogue contexts (Adiwardana et al., 2020; Roller et al., 2021; Zhang et al., 2020). However, there is room for further improvement even for recent RGSs. For example, even the responses generated by recent RGSs sometimes generate inappropriate responses in a dialogue with the user, such as responses based on incorrect knowledge or responses containing offensive expressions (Shuster et al., 2022). Among various issues, contradictory responses pose a particularly grave concern. A contradiction not only disrupts the dialogue flow but also creates a detrimental perception of the RGS lacking comprehension of the dialogue content (Li et al., 2022; Nie et al., 2021). Moreover, as described in Chapter 2, contradictions influence the occurrence of other errors. Effectively improving RGSs' ability to avoid contradictory responses, Consistency-Awareness (CA), is crucial in developing RGSs that can establish a trustworthy and symbiotic relationship with users.

Automatic evaluation for the ability to avoid contradictory responses of RGSs has the potential for the improvement of CA because of its high reproducibility and low cost. Specifically, CA can be improved efficiently by repeatedly improving RGSs using automatic evaluation with a certain level of accuracy and low cost, and using high-cost human evaluation only for the final validation. The development of an automatic evaluation framework that is reliable, i.e., highly correlated with human CA evaluations, is critical to effectively improving the CA of RGSs.

A straightforward method of automatic CA evaluation is to calculate the frequency of RGS-generated contradictory response generation using a contradiction detector that binary classifies whether a response contains a contradiction (Nie et al., 2021; Welleck et al., 2019). We call this approach generation-based automatic evaluation. However, this method requires a contradiction detector capable of detecting contradictory responses with high accuracy. As shown in Chapter 4, the accuracy of the current best-performance detector is as low as 0.54 for binary classification, making practical automatic generation-based evaluation difficult. Therefore, in recent years, automatic CA evaluation based on the assignment of generation probabilities to responses that have been prepared in advance is often employed as an alternative approach (Kim et al., 2020; Welleck et al., 2019). Specifically, when a noncontradictory or contradictory response is prepared for a certain dialogue context, and the generation probabilities of these responses are calculated by the evaluation target RGS, the RGS is evaluated based on whether it can assign a high generation probability to the noncontradictory response or a low generation probability to the contradictory response. We call this method probability-based automatic evaluation. The advantage of probability-based automatic evaluation is that once a noncontradictory or contradictory response is prepared, the RGS can be evaluated without automatically evaluating the responses generated by the RGS, thus allowing automatic evaluation of CAs without a high-performance contradiction detector, which is currently not available. However, it has not been fully verified whether probability-based automatic evaluation can really be the alternative evaluation of generation-based evaluation. That is, no earlier studies validate the correlation between the results of probability-based automatic evaluation and whether or not RGS actually generates contradictory responses.

In this study, we explore and construct a framework for highly effective and practical automatic evaluation of CA. First, we confirm that highly analytical automatic evaluation of RGS is difficult with probability-based automatic methods. Specifically, we confirm through experiments that there is no correlation between the results of probability-based automatic evaluation and those of human evaluation at the instance level, which is a highly finer and interpretable evaluation level. We then undertake two tasks to improve the effectiveness and practicality of generation-based automatic evaluation of CA. First, in order to improve the effectiveness of automatic generation-based evaluation, we improve the performance of contradiction detectors by augmenting their training data. The training data is not a set of human-generated contradictory responses, as in the past, but actual contradictory responses generated by RGS to reduce the gap between the contradiction detector training data and the inference target. Second, in addition to addressing the existing challenge of improving the accuracy of the inconsistency detector, we develop a practical generation-based framework for automatically evaluating the CA of RGSs, assuming their deployment. A practical

method to avoid contradictory responses from RGS is to use the contradiction detector in post-processing to select noncontradictory response candidates (Nie et al., 2021; Welleck et al., 2019). Considering such post-processing, the final output of the RGS may depend not only on the consistency of the final output determined by RGSs themselves but also on whether multiple RGS response candidates contain noncontradictory responses. Therefore, we propose a method for generation-based automatic CA evaluation that takes into account response candidates generated by RGSs.

1.1 Research Issues

This thesis addresses the following research issues:

- **Which approach should we employ for automatic CA evaluation?** Although probability-based automatic evaluation allows the evaluation of CAs without a contradiction detector unlike generation-based ones, the validity has not been tested.
- **Can we realize effective generation-based CA evaluation?** The accuracy of the current best-performance detector is not high enough for practical automatic generation-based evaluation. For effective CA evaluation, their performance needs to be improved.
- **Can we realize practical generation-based CA evaluation?** In addition, considering post-processing using a contradiction detector, the consistency of all the candidates in n-best generated by RGS is also important for CA evaluation. Nevertheless, conventional generation-based automatic evaluation focuses only on the consistency of the 1-best candidates generated and does not fully analyze the characteristics of the n-bests generated by RGS.

1.2 Contributions

This thesis makes the following contributions:

- **Confirmed that we need to employ generation-based evaluation for efficient RGS improvement.** We experimentally confirmed through experiments that there is no correlation between the results of probability-based automatic evaluation and those of human evaluation at the instance level.

- **Improved the accuracy of data-driven contradiction detectors.** We collected a large collection of RGS-generated contradictory responses for training data-driven contradiction detectors. Our experiments demonstrated that training detectors on our dataset improved the accuracy of contradiction identification.
- **Proposed an n-best-aware CA evaluation framework.** We propose evaluating CA considering the consistency of n-best candidates generated by RGSs, assuming the post-processing where an ideal contradiction detector chooses noncontradictory candidates from the n-best candidates.

1.3 Thesis Overview

An overview of this paper is given as follows:

Chapter 2. Background. We introduce the background of automatic CA evaluation.

Chapter 3. Identifying effective approach for automatic CA evaluation. Our experiments demonstrate that probability-based automatic CA evaluation has no correlation with human CA evaluation in instance-level evaluation, which is essential for efficient CA improvement.

Chapter 4. Improving automatic contradiction detector. We report our large-scale data collection for augmenting the training resource of automatic contradiction detectors, along with the description of our experiments in which we confirmed that training detectors with our collection improves contradiction identification accuracy.

Chapter 5. Expanding evaluation target to n-best responses. We propose a framework for automatic evaluation of CA considering the consistency of their n-best candidates assuming the post-processing where an ideal contradiction detector chooses noncontradictory candidates from the n-best candidates.

Chapter 6. Conclusions. We summarize our contributions to realize effective automatic CA evaluation.

Chapter 2

Background

2.1 Neural Dialogue Response Generation Systems

Recent RGS advances. The exploration of dialogue response generation systems (RGS) has captured widespread attention across various domains, including medicine and education (Addlesee et al., 2019; Litman et al., 2016), and is currently a focal point of active research and development endeavors. Notably, the emergence of deep learning-based RGS has undergone rapid exploration in recent years, promoted by advancements in deep learning technology. These systems are known to be able to generate fluent responses to dialogue contexts (Adiwardana et al., 2020; Roller et al., 2021; Zhang et al., 2020). In particular, it has been reported that the evaluation results of engagingness, the degree to which a user is willing to talk to the agent, are comparable to that of human beings (Roller et al., 2021).

Errors of current RGS. Despite these advancements, there is room for further enhancement in the pursuit of RGSs capable of engaging in natural conversations with humans. Even the responses generated by recent RGSs occasionally exhibit errors that would not occur in human conversation. Among the various types of reported inappropriate responses, the major domain-independent errors include context-irrelevant responses, contradictory responses, factually incorrect responses, and offensive responses (Kann et al., 2022; Shuster et al., 2022), as shown in Figure 2.1 These errors are known to decrease users' willingness for human-bot interaction, resulting in the breakdown of dialogue (Martinovsky and Traum, 2003). These outcomes are especially critical in RGS applications that interact with users over long periods of time, such as medical and educational applications. To achieve a reliable and engaging RGS, these errors must be suppressed.

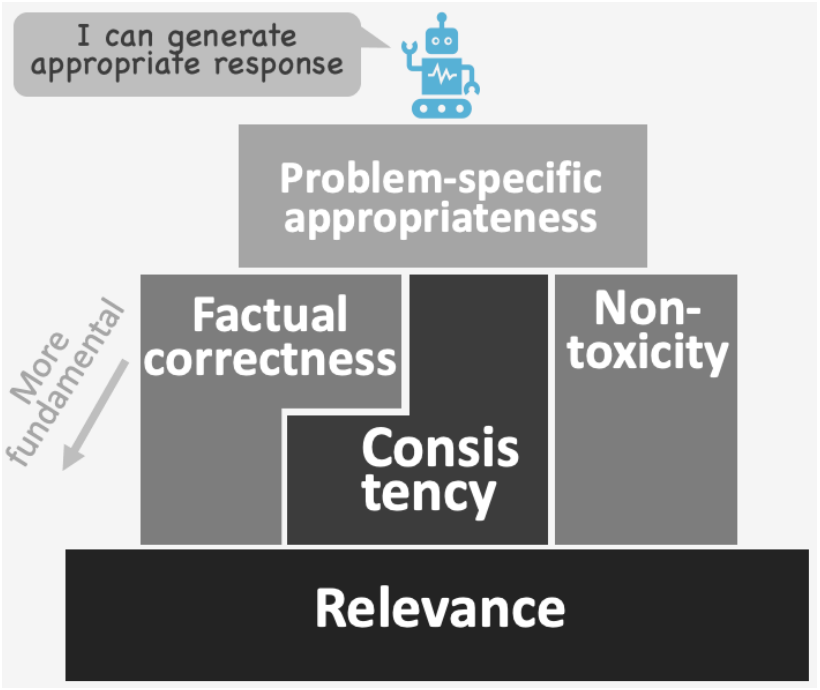

Figure 2.1 Representative semantical errors of RGSs. In addition to domain-independent errors, developers need to improve their RGSs to guarantee problem-specific appropriateness.

2.2 Contradiction Awareness

2.2.1 Contradiction types

Two major contradiction types are identified in the context of dialogue response generation: (i) contradictions against the facts in the world outside of the ongoing dialogue (e.g., personas) and (ii) those against what is stated in the local preceding context (e.g., opinions) (Li et al., 2020; Nie et al., 2021). This study focuses on suppressing the second type. While several studies have addressed the issue of avoiding the first contradiction type (Kim et al., 2020; Kottur et al., 2017; Li et al., 2016; Qian et al., 2018; Zhang et al., 2018), given that the multi-turn human-bot interaction is attracting increasing interest, we believe that tackling the issue of the second type is becoming increasingly important.

2.2.2 Importance of consistency awareness

As noted in the above section, RGS has a wide variety of errors. Among them, the generation of context-irrelevant responses has been solved to a certain extent by companies with large computational and human resources, which have been scaling up their deep learning

models (Adiwardana et al., 2020; Zhang et al., 2020). Therefore, contradictory responses, factually incorrect responses, and offensive responses are the representative front-line errors that have not been resolved. Out of these three error types, we believe it is especially important to prioritize the suppression of contradictions. There are two major reasons for this focus. The first reason is that contradiction itself has a serious impact on the dialogue, as described in Chapter 1. Another reason is that the suppression of contradictions also plays an important role in suppressing the other two errors, as described in the following paragraphs.

Relations with suppressing factually incorrect responses. Factually incorrect responses are inconsistent with facts in the world. One effective suppression method for factually incorrect responses is referring to factually correct external documents (Ji et al., 2023). However, the current situation is far from resolved. For example, not even a widely accepted automatic evaluation method has been established (Ji et al., 2023). Therefore, addressing this error must obviously be studied. Here, it has been suggested that some of the factually incorrect responses may be suppressed by improving the consistency of the RGS (Mündler et al., 2023). Thus, consistency awareness needs to be improved first to address factual errors, at least in part.

Relations with suppressing offensive responses. One effective suppression method for offensive responses is training RGSs on human-written examples of ideal agent outputs (OpenAI, 2023; Ouyang et al., 2022). This type of training is tackled by companies with large budgets since it requires large-scale data construction costs. Despite these efforts, the existing circumstances remain considerably unresolved. For instance, context-aware toxicity handling has not been fully explored. Thus, suppressing offensive responses also needs to be studied. Although there is no strong direct relationship between suppressing contradictions and offensive responses, it has been reported that some of the techniques used to suppress contradictions are also effective in addressing toxicity (Goldzycher et al., 2023). Therefore, brewing technologies for the suppression of contradictions may, in part, also help to suppress the generation of offensive responses.

2.2.3 Consistency awareness of current RGSs

As mentioned above, the suppression of contradictions is essential, and various efforts have been made in previous studies. The mainstream approaches of prior studies to address contradictions have been data-driven. Welleck et al. (2019) developed a dialogue-domain natural language inference dataset by applying a rule-based method to transform an existing dialogue corpus. They employed this dataset to train a contradiction detector that automatically

identifies contradictions within pairs of dialogue domain sentences. Nie et al. (2021) gathered and employed contradictory and noncontradictory human-written responses to train a contradiction detector. Meanwhile, Li et al. (2020) and Li et al. (2022) updated RGMs using a loss function that reduces the likelihood of generating inconsistent responses to suppress contradictions.

Despite these attempts, the suppression of contradictory responses remains far from resolved. There are even cases of contradictions with their own utterances immediately before the responses, not to mention contradictions with utterances in past dialogues (Chapter 4).

Chapter 3

Identifying effective approach for automatic CA evaluation

3.1 Introduction

Improving CA is an important issue for RGS, and several previous studies have attempted to realize automatic CA evaluation. At present, evaluation methods can be divided into two types of approaches: generation-based (Nie et al., 2021; Welleck et al., 2019) and probability-based (Kim et al., 2020; Welleck et al., 2019), as described in Chapter 1. Generation-based evaluation is an approach to reproduce human evaluation, but it has not been put into practical use due to the low performance of existing contradiction detectors (Chapter 4). Due to this background, probability-based evaluation has been attracting attention as a substitute for generative-based evaluation. Probability-based evaluation has been actively used in recent years due to its advantage of not requiring an automatic evaluation system such as a contradiction detector for generation-based (Kim et al., 2020; Welleck et al., 2019). However, the effectiveness of automatic evaluation based on this approach has not yet been verified, as described in Chapter 1. If probability-based evaluations cannot evaluate CA effectively, CA improvements guided by this evaluation may actually go in the wrong direction and may significantly hinder CA improvement. Therefore, it is extremely important to verify the validity of the probability-based evaluation, or in other words, to verify whether it is necessary to evaluate CAs with generation-based evaluation although it requires an accurate automatic evaluation system.

In this study, we experimentally test the effectiveness of probability-based automatic evaluation. In general, not only for CA, there are two levels of granularity in RGS evaluation: system-level and instance-level (Lowe et al., 2017). As discussed in the following

section, the ability to evaluate at either granularity is important, but in particular, instance-level evaluation, which is a more granular evaluation, plays an important role in improving the CA. Therefore, prior to using probability-based automatic evaluation as a reliable framework for automatic CA evaluation, it is necessary to verify whether automatic CA evaluation correlates with human evaluation not only at the system level but also at the instance level. Thus, we examine whether the results of probability-based automatic evaluation correlate with the results of human evaluation at both the system and instance levels of granularity. In order to gain diverse insights, we examine not only the evaluation of CA, but also a more general aspect of CA: the evaluation of response appropriateness. As a specific method of probability-based automatic evaluation, we employ response selection, which is one of the representative probability-based evaluation methods (Kim et al., 2020; Welleck et al., 2019). The results of the validation with response selection show that the probability-based automatic evaluation results have a certain correlation with the human evaluation at the instance level, but do not correlate with the human evaluation at the instance level, for both CA and appropriateness. Therefore, it is necessary to improve and employ generation-based evaluation for highly effective automatic evaluation of CA and appropriateness.

3.2 Evaluation Levels

Not only for CA (or appropriateness), there are two levels of granularity in RGS evaluations: system-level and instance-level (Lowe et al., 2017). The system level is more coarser, while the instance level is more granular. In this study, we test whether the results of probability-based automatic evaluations of CA and appropriateness correlate with the results of human evaluations at two granular levels.

3.2.1 System-level evaluation

System-level evaluation refers to an evaluation in which, given multiple RGSs to be evaluated, the performance of the RGSs is compared. For instance, in the example at the top of Figure 3.1, given four RGSs, a ranking is created based on the results of the system-level evaluation. System-level evaluation mainly helps to perform comparisons between pre- and post-improvement RGSs.

3.2.2 Instance-level evaluation

The instance level, on the other hand, refers to an evaluation that predicts the quality of the target RGS's response to a specific dialogue context. For instance, in the example at

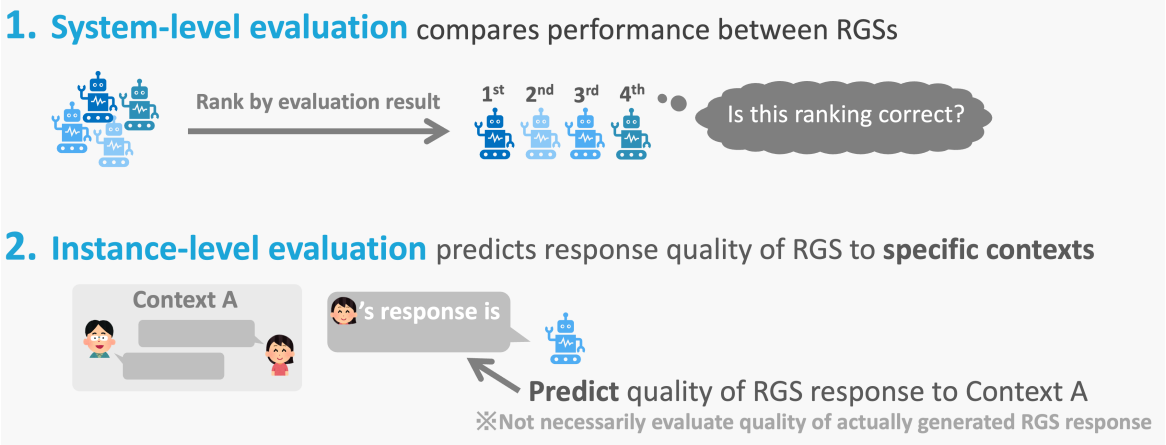

Figure 3.1 Two evaluation levels: (1) system-level and (2) instance-level.

the bottom of Figure 3.1, given a dialogue context, Context A, it predicts the quality of the response that the target RGS will generate to Context A. Instance-level evaluation has two characteristics. First, instance-level evaluation can provide important analysis for efficient RGS improvement, such as in which dialogue contexts the RGS has trouble generating responses. Second, if instance-level evaluation is possible, system-level evaluation is also possible based on the percentage of dialogue contexts in which RGS has problems generating responses. Given these characteristics, it is important to have an automatic evaluation framework that can perform highly effective instance-level evaluations.

Note that instance-level evaluation "predicts" the quality of the RGS response to a particular context, not necessarily evaluating the quality of the "generated" response. For example, probability-based evaluation does not evaluate the quality of responses generated by the RGS for a given dialogue context but rather predicts the quality via observing the behavior of the RGS for the context based on probability assignments for prepared responses. These probability-based evaluations are also included in instance-level evaluations.

3.3 Dialogue Response Selection

In this study, we verify the effectiveness of probability-based automatic evaluation of CA and appropriateness. For the verification, it is necessary to prepare a specific method for probability-based automatic evaluation. In this study, we use dialogue response selection as a representative method of probability-based automatic evaluations (Kim et al., 2020; Welleck et al., 2019).

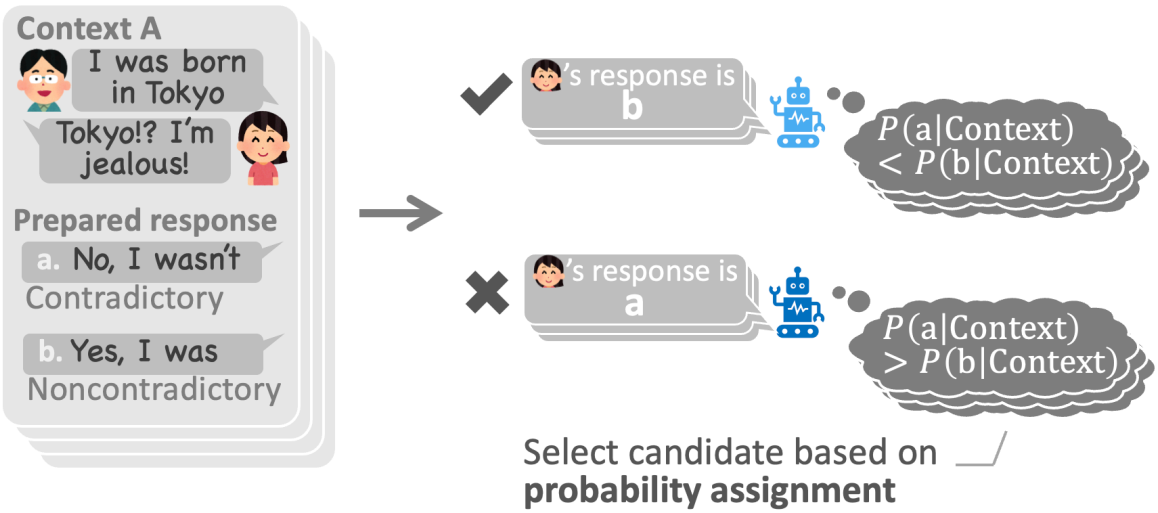

Figure 3.2 An example of dialogue response selection for CA evaluation.

Dialogue response selection is the task of selecting an appropriate candidate given a dialogue context and subsequent response candidates. Each question of the dialogue response selection task for CA evaluation consists of a dialogue context, a correct candidate that is an appropriate and noncontradictory response to that context, and an incorrect candidate that is a contradictory response to that context. In the case of evaluating appropriateness, a correct candidate should be an appropriate response, and an incorrect candidate should have at least one semantic error as a response to that context. RGSs are required to select the correct candidate.

In this section, we describe the way to have RGSs solve for dialogue response selection and the construction of the dialogue response selection dataset used in this validation.

3.3.1 Response selection by RGSs

Basically, RGSs are not designed to select response candidates directly. Instead, RGSs compute the softmax cross-entropy loss ℓ_r for each response candidate $r \in \mathscr{R}$. The candidates with the lowest losses are regarded as the RGSs's selections: $\hat{r} = \underset{r \in \mathscr{R}}{\operatorname{argmin}} \ell_r$. Figure 3.2 shows an example of dialogue response selection for CA evaluation.

3.3.2 Response selection dataset construction for evaluating appropriateness

In this section, we describe the dataset construction method of dialogue response selection and the results of the construction for appropriateness evaluation.

Method for dataset construction

For each context c and ground-truth response r^{true}, we construct a set of false response candidates $r^{\text{false}} \in \mathscr{R}^{\text{false}}$ by gathering utterances from an utterance repository $u \in \mathscr{U}$.

We collect responses that are semantically inappropriate to serve as false candidates for the response selection test set, allowing us to assess RGSs' sensitivity to appropriateness. Simply extracting responses at random from the repository might lead to the inclusion of predominantly negative candidates that bear little relevance to the dialogue context. Consequently, the evaluation of the RGS is constrained to measuring its ability to recognize the topic relevance between dialogue contexts and their response candidates. To broaden the assessment of RGS sensitivity to various types of inappropriateness, it becomes imperative to gather false candidates displaying a diverse range of inappropriate characteristics. In this study, we employ the following method to collect utterances that, while not entirely unrelated to the dialogue context, are deemed inappropriate as responses:

1. Retrieve M utterances, $\{u_1, \cdots, u_M\}$, related to the ground-truth response r^{true} from the utterance repository $\mathscr{U}$.
2. Remove acceptable ones from the retrieved utterances by human evaluation.

1. Retrieve utterances related to the ground-truth response. We assume that utterances related to the ground-truth response share some similar content words between them. Here, we retrieve the related utterances on the basis of the similarities of the content words. To make false candidates in each pool diverse, we use two retrieval methods: lexical retrieval and embedding-based retrieval. We use Lucene[1] for lexical retrieval, and cosine similarity of sentence vectors for embedding-based retrieval. Sentence vectors are SIF (Arora et al., 2017) weighted average of ELMo word vectors (Peters et al., 2018).

2. Remove acceptable utterances. Coincidentally, some of the retrieved utterances may be acceptable as an appropriate response. To remove such utterances, we ask human annotators to evaluate each retrieved utterance. Specifically, we instruct five annotators (per

[1] https://lucene.apache.org/.

candidate) to score each retrieved candidate in a five-point scale from 1 to 5. A score of 5 means that the utterance can clearly be regarded as an appropriate response for the given context, whereas a score of 1 means that it cannot be regarded as an appropriate one at all. In addition to the scores, we also instruct annotators to give a score of 0 to ungrammatical utterances. We remove the utterances that are given a score of 3 or higher by three or more annotators because these utterances with a high score can be acceptable. In addition, we remove the utterances that are given a score of 0 by three or more annotators because these are likely to be ungrammatical ones. We also instruct annotators to score ground-truth responses, combining them with retrieved utterances. We remove the questions if the score of the ground-truth response is low, i.e., three or more annotators give a score of 3 or lower. This is intended to ensure that ground-truth responses are certainly appropriate for the given context.

Results of dataset construction

Settings of test set construction. We retrieve 10 utterances (per question) from the repository and remove acceptable ones following the method described in Section 3.3.2. We use crowdsourcing[2] to score the retrieved utterances. After removing acceptable utterances, there are some questions that have 6 or more available false candidates. From these questions, we develop new questions with the same context but different candidates (both ground-truth responses and false candidates). We regard one of acceptable utterances removed by human evaluation as the ground-truth responses of new questions. We use the dialogue data from DailyDialog (Li et al., 2017) to construct the test set. We extract the four beginning turns of each dialogue sample from DailyDialog, regarding the fourth utterance as the ground-truth response. We extract the utterances of OpenSubtitles2018 (Lison et al., 2018) to construct the repository used to retrieve false candidates. Note that the repository does not contain the utterances in the dialogue data used to train response generation systems in all subsequent experiments.

Statistics of our test set. We developed the test set that consists of $1,019$ questions with 4 candidates (1 ground-truth + 3 false candidates). Table 3.1 shows the basic statistics of our test set. The Fleiss' Kappa of the annotators' scoring in the six scale is 0.22.[3] Note that if we regard the scoring as binary classification (scores higher than 3 are regarded as appropriate responses, and the others not), the Fleiss' Kappa of the scoring is 0.63, which seems to be reasonably high.

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

[3]We calculated Fleiss' Kappa based on the scale of the scores as categorical.

Table 3.1 Basic statistics of our test set.

Total questions	1,019
Candidates per question	4
Context turns per question	3
Kappa of the scoring (six classes)	0.22
Kappa of the scoring (two classes)	0.63

Table 3.2 Example of our test set. All three false candidates contain the content word "focus", which is related to the context (topic).

Context:	
A:	Excuse me. Could you please take a picture of us with this **camera?**
B:	Sure. Which button do I press to shoot?
A:	This one.
Candidates:	
1.	Could he not **focus** on that?
2.	But I do have ninja **focus**.
3.	Do not lose your **focus**!
4.	Do I have to **focus** it? [Ground-truth]

Example of our test set. Table 3.2 shows an example of our test set. All the false response candidates share the same content word "focus" related to the topic "camera".

Preliminary experiments. We conducted a simple experiment to investigate whether or not a system that takes only content words into account can recognize false response candidates in our test set. For the model, we used the TF-IDF model (Lowe et al., 2015), which simply compares between content words of a given context and each candidate. As a result, the accuracy was 0.461. For a comparison, we also replaced all the false candidates in our test set with randomly sampled utterances. The accuracy of the same TF-IDF model increased to 0.671. These results indicate that we have successfully collected false candidates that are not irrelevant but inappropriate.

3.3.3 Response selection dataset construction for evaluating CA

From the test set collected in the previous section, we construct a response selection test set for CA evaluation. Specifically, among the false candidates in the test set for appropriateness evaluation, we extract only those determined to be inappropriate due to inconsistency with

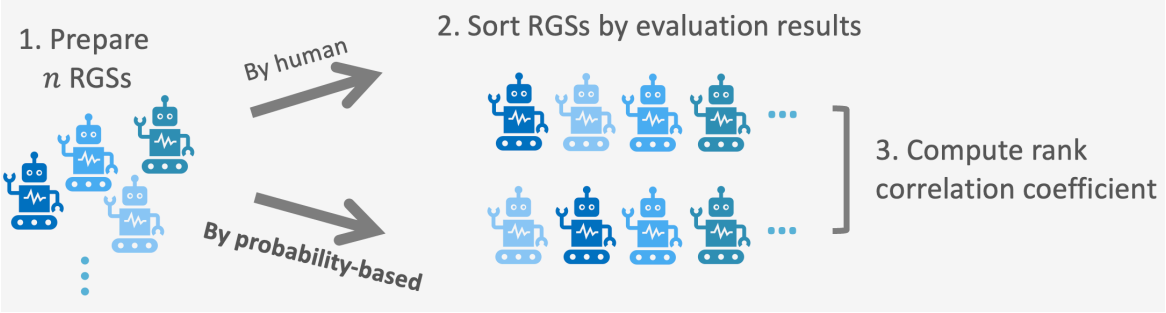

Figure 3.3 Procedures of the validation for system-level evaluation.

the dialogue context and use them as new false candidates of a response selection test for CA evaluation.

Method for dataset construction

From the false candidates of the test set developed in Section 3.3.2, we collect false candidates contradicting their contexts through human annotation. Specifically, each of the false candidates is classified into three classes by three human annotators: contradictory to dialogue context, irrelevant to dialogue context, or containing other errors. We retrieve candidates classified in the first class above by two or more annotators as candidates containing contradictions with their contexts.

Results of dataset construction

Settings of test set construction. We employed 692 false candidates of the response selection test set for evaluating appropriateness. We use crowdsourcing[4] to score the retrieved utterances.

Statistics of our test set. We developed the test set that consists of 50 questions with 2 candidates (1 ground-truth + 1 false candidates).

3.4 Validation for System-level Evaluation

First, we investigate whether the results of probability-based automatic evaluation of CA and appropriateness correlate with the results of human evaluation at a rougher granular system level.

[4]https://www.mturk.com/.

3.4.1 Procedures

Figure 3.3 shows the validation method. This consists of ranking n RGSs in each of the probability-based-automatic and manual evaluations, and calculating the rank correlation coefficient between the two rankings.

Rank by probability-based automatic evaluation. As described in the previous section, we employ the response selection task as a probability-based automatic evaluation method. For each of the evaluation target RGSs, we calculate the percentage of questions among m response selection questions in which the RGS selects ground-truth candidates. The ranking of the RGSs is based on this percentage.

Rank by human evaluation for appropriateness. Evaluation target RGSs generate a response r^{gen} for each input context $c \in \mathscr{C}$ used in the response selection. Then, five human annotators (per response) score each generated response r^{gen} in a five-point scale from 1 to 5. A score of 5 means that the response can clearly be regarded as an appropriate response for the given context, whereas a score of 1 means that it cannot be regarded as an appropriate one at all. As a result, we obtain five scores, $\{s_1, s_2, \cdots, s_5\}$, for each response r^{gen} and average them: $s^{\mathrm{mean}} = \mathrm{mean}(s_1, s_2, \cdots, s_5)$. We also average s^{mean} across all the questions in the test set and yield the final score s^{final} for each RGS. Based on this score, we make a ranking of the RGSs. Although we developed the test set that consists of $1,019$ questions, it is too costly to evaluate several RGSs' responses for $1,019$ questions by humans. Thus we give the context of 50 randomly sampled questions from our test set to evaluation target RGSs as inputs $\mathscr{C}$.

Rank by human evaluation for CA. Evaluation target RGSs generate a response r^{gen} for each input context $c \in \mathscr{C}$ used in the response selection. Then, three human annotators (per response) classify each generated response r^{gen} into two classes: contradictory or noncontradictory. A response judged to be contradictory by two or more annotators is regarded as a contradictory response. We compute the percentage of noncontradictory ones among all generated responses for each RGS. Based on this percentage, we make a ranking of the RGSs. We employ all 50 contexts included in the response selection test set for CA as inputs $\mathscr{C}$.

Evaluation target RGSs. In order to avoid experimental results dependent on specific RGSs, 20 diverse RGSs were prepared. Table 3.4 lists all employed RGSs.

Table 3.3 Spearman's rank correlation coefficient between ranking tables created by manual evaluation and those created by automatic evaluation.

Perspective	Spearman	p-value
Appropriateness	0.74	$< 1.0 \times 10^{-2}$
CA	0.60	$< 1.0 \times 10^{-2}$

3.4.2 Results

Table 3.3 displays Spearman's rank correlation coefficient between the ranking created by human evaluation and automatic evaluation for the two perspectives. The rank correlation coefficients for both perspectives reached 0.6, indicating that the results of automatic evaluation by response selection have a certain degree of correlation with the results of manual evaluation at the system level. These results suggest that at the system level, probability-based automatic evaluation may be able to assess the CA and appropriateness of RGS with reasonable effectiveness.

3.5 Validation for Instance-level Evaluation

Second, we investigate whether the results of probability-based CA and appropriateness evaluation correlate with the human evaluation results at a finer granular instance level.

3.5.1 Procedures

For the above system-level validation, a dialogue response selection test set was constructed for each of the appropriateness and CA aspects, and 20 RGSs solved them. We then extracted the dialogue contexts from the two test sets and manually evaluated the responses generated by the same 20 RGSs in terms of appropriateness and CA. Therefore, we can analyze the instance-level correlation of probability-based evaluation by comparing an RGS's response selection result for a dialogue context used for human evaluation, with the human evaluation result of the same RGS for the corresponding dialogue context.

Appropriateness. If an RGS can generate highly appropriate responses in dialogue contexts where the RGS can select ground-truth in response selection, then the human evaluation results for the RGS responses to dialogue contexts in which the RGS correctly selects a candidate in response selection should score higher. We therefore divide the 50 dialogue contexts used for human evaluation into two groups, depending on whether the RGS was correct in its

response selection for each of the 20 RGSs. We then examine for each RGS whether there is a difference in the average of the manual ratings of the appropriatenss between the two groups.

CA. The human evaluation of CA at the system level is a binary classification, not a rating like the evaluation for appropriateness. Therefore, the average of the human ratings cannot be compared between two groups classified based on the results of response selection. Instead, we compare the contradiction frequency between the two groups.

3.5.2 Results

Appropriateness. For each of the 20 RGSs, we compared the averages of the human ratings between two dialogue context groups based on the response selection results, using a one-tailed t-test at 5% significance level. The results showed that for only 4 out of 20 RGSs, the average human evaluation scores of the response-selection-correct group were statistically significantly higher than those of the response-selection-incorrect group.

CA. For each of the 20 RGSs, we compared the number of human-judged noncontradictory responses between two dialogue context groups based on the response selection results using Fisher's exact test at 5% significance level. The results showed that for none of 20 RGSs, the number of contradictory responses of the response-selection-correct group were statistically significantly higher than those of the response-selection-incorrect group.

These results indicate that for instance-level RGS evaluations, the probability-based automatic evaluation results for appropriateness and CA do not necessarily correlate with the results of the human evaluation.

3.6 Evaluation considering n-best

In the previous section, we found that at the instance level, there is not necessarily a correlation between the human evaluation of RGS's 1-best response and the evaluation results from response selection. We finally investigate whether this result is also true for the correlation with the human CA evaluation of the n-best RGS responses, that is, the evaluation of the percentage of contradictory responses in generated n-bests. As discussed in Chapter 5, considering n-best, not only 1-best, is essential for avoiding contradictions of RGS. Therefore,

it is important to be able to automatically evaluate the appearance of contradictions within RGS's n-best to avoid contradictory outputs in practice.

3.6.1 Procedures

Each evaluation target RGS generates n-best responses for each input context. We evaluate the consistency of each response candidate in the generated n-best list and calculate the contradictory responses' percentage of the n-best list, which we treat as a score of the n-best list. We compare this n-best score with the evaluation result through response selection.

Due to the high cost of evaluating the n-best of all 20 RGSs used in the experiments up to the previous section, we evaluate the n-best of 10 of these RGSs.[5] For each of the 10 RGSs, we manually obtained the percentages of contradictory responses for a total of 10 n-bests by taking the 5 dialogue contexts answered correctly and the 5 dialogue contexts answered incorrectly in the dialogue response selection for CA evaluation of the corresponding RGS. We finally compare all 10 RGSs' average contradictory response percentages of the two dialogue context groups classified based on dialogue response selection results. The value of n was set to 10.

3.6.2 Results

The mean percentage of contradictory responses in the 10-best of the 10 RGSs in the response-selection-correct group was 2.6. On the other hand, the mean percentage in the response-selection-incorrect group was 2.5. This indicates that RGS does not necessarily generate n-bests with a higher proportion of contradictory responses for dialogue contexts where the RGS incorrectly answers in response selection. The proportion of contradictory responses in n-bests, i.e., the ability of RGS to avoid contradictory responses through generating n-bests, cannot necessarily be evaluated by probability-based CA evaluation.

3.7 Conclusion

There are two approaches to CA evaluation: generation-based and probability-based. The probability-based approach has the advantage of allowing evaluations without high-performance contradiction detectors, but the effectiveness of such evaluations has not been verified.

[5] We employed RGS #1 through #10 in Table 3.4.

In this study, we examined the effectiveness of probability-based automatic CA evaluation by measuring the correlation between human evaluation and automatic evaluation with response selection at two levels: the system level and the instance level. At the system level, the results of automatic evaluation by response selection were found to correlate to a certain degree with the results of human evaluation, not only for CA but also for appropriateness. On the other hand, at the instance level, both CA and appropriateness had no correlation with human evaluation in our settings. From the above, it was found that automatic evaluation of CA requires the use of a generation-based approach, which evaluates the responses actually generated by RGSs.

Table 3.4 Training data and hyperparameters of evaluation target RGSs.

	Architecture	Training data	# of Ec/Dc[*1]	# of hidden dim[*2]	# of emb dim[*3]	# of input[*4]
1	GRU	OpenSub[*5]	1/ 1	256	256	3
2	GRU	OpenSub	1/ 1	512	512	3
3	GRU	OpenSub	2/ 2	256	256	3
4	GRU	OpenSub	2/ 2	512	512	3
5	LSTM	OpenSub	1/ 1	256	256	3
6	LSTM	OpenSub	1/ 1	512	512	3
7	LSTM	OpenSub	2/ 2	512	512	1
8	LSTM	OpenSub	2/ 2	512	512	3
9	LSTM	OpenSub→Self[*6]	2/ 2	512	512	3
10	ConvS2S	OpenSub	20/20	3×512	512	3
11	ConvS2S	OpenSub→Self	20/20	3×512	512	3
12	Transformer	OpenSub	2/ 2	4[*7]	256	3
13	Transformer	OpenSub	2/ 2	4[*7]	512	3
14	Transformer	OpenSub	4/ 4	4[*7]	256	3
15	Transformer	OpenSub	4/ 4	4[*7]	512	3
16	DialoGPT-large	-[*8]	-/36	20[*7]	1280	3
17	DialoGPT-medium	-[*8]	-/24	16[*7]	1024	3
18	DialoGPT-medium	Self[*9]	-/24	16[*7]	1024	3
19	DialoGPT-small	-[*8]	-/12	12[*7]	768	3
20	DialoGPT-small	Self[*9]	-/12	12[*7]	768	3

[*1] These indicate the number of layers of encoder/decoder.
[*2] These indicate the number of hidden layer's dimensions.
[*3] These indicate the number of embedding's dimensions.
[*4] These indicate the number of utterances used as context.
[*5] Training was performed on 5M dialogue pairs constructed using OpenSubtitles2018.
[*6] The training was performed on 5M dialogue pairs constructed using OpenSubtitles2018, and then additional training was performed on 0.27M dialogue pairs constructed using Self-dialogue Corpus (Fainberg et al., 2018; Krause et al., 2017).
[*7] These indicate the number of attention heads.
[*8] https://github.com/microsoft/DialoGPT.
[*9] Additional training was performed on dialogue pair data constructed using Self-dialogue Corpus, using publicly available model parameters as initial parameters.

Chapter 4

Improving automatic contradiction detector

4.1 Introduction

It was shown in Chapter 3 that for highly effective automatic CA evaluation, it is necessary to employ generation-based automatic evaluation methods. In automatic generation-based evaluation, a contradiction detector plays a very vital role. The contradiction detector is required to perform binary classification of whether a given RGS response contains contradictions or not. It is no exaggeration to say that the effectiveness of automatic generation-based evaluation depends on the accuracy of the detector's contradiction identification.

Previous studies demonstrated the effectiveness of data-driven methods in developing contradiction detectors by utilizing datasets containing contradiction instances (Nie et al., 2021; Welleck et al., 2019). As evidenced in various natural language processing tasks (Leite et al., 2020; Mosbach et al., 2020), the data-driven system performance is expected to be contingent upon the accessible data volume. Therefore, the performance of data-driven contradiction detectors can be enhanced if a large number of contradictory responses from RGSs are available. However, a comprehensive collection of RGS-generated contradiction data is lacking, except for the small dataset of less than 400 RGS-generated responses compiled by Nie et al. (2021), which is intended for use as test data. Consequently, data-driven developments have resorted to alternative sources, such as automatically synthesized or human-written contradictions (Li et al., 2022; Nie et al., 2021). These substitute responses possess different characteristics compared to RGS-generated ones (Section 4.5.1), causing a gap between the training data and the inference targets in practice. Therefore, the full potential of data-driven methods may not be realized.

Table 4.1 Dataset examples in which the speakers are identified as A and B. CG, B1, and BL are the three distinct RGSs. The labels provided by three human workers are represented by C and N, which denote contradictory and noncontradictory responses, respectively. The bolded portion illustrates a contradiction.

Context
A: **I hurt my toe doing ballet.** — u_n
B: Oh I hope you get better. Does is hurt a lot?
A: It hurts pretty bad, but it will heal. [. . .]
B: [. . .] Do you do ballet practice often? — u_q
RGS responses on speaker A's side
CG: **I don't do ballet myself,** I was just watching a performance. [. . .] — [C, C, C]
B1: **I have never done ballet,** but I love the music. I listen to it all the time. — [C, C, C]
⋮
BL: Yes, I do ballet every day. [. . .] — [N, N, N]

In this study, we demonstrate the effectiveness of having access to a vast repository of RGS-generated contradictory responses in tackling the development of contradiction detectors. We first collect a substantial amount of contradictory responses from various high-performance RGSs. An example of our dataset is shown in Table 4.1. We then demonstrate that a contradiction detector trained on human-written contradiction data exhibits limited accuracy in identifying RGS contradictions and confirm that training on our dataset improves the situation. We also comprehensively analyze our collection from various angles, identifying differences in characteristics between the RGS and human contradictory responses.

4.2 Related studies

4.2.1 Improvement of contradiction detectors with contradictory responses

The mainstream approaches of prior studies to develop contradiction detectors have been data-driven. Welleck et al. (2019) developed a dialogue-domain natural language inference dataset by applying a rule-based method to transform an existing dialogue corpus. They employed this dataset to train a contradiction detector that automatically identifies contradictions within pairs of dialogue domain sentences. Nie et al. (2021) gathered and employed

human-written contradictory responses to train a contradiction detector. The efficacy of these data-driven contradiction detectors could further be enhanced by having access to various RGS-generated contradictory responses because they need to deal with RGS-generated contradictions practically. The present study aims to collect these contradictions from RGSs to provide valuable resources for a wide range of data-driven contradiction suppression methods including the improvement of contradiction detectors. To the best of our knowledge, this is the first work to gather a substantial volume of contradictory responses from RGSs for training data development.

4.2.2 Effective inputs to collect contradictions

Previous studies demonstrated that RGSs tend to generate contradictory responses when previously stated facts or opinions are repeated (Li et al., 2021; Nie et al., 2021). Nevertheless, posing questions that prompt dialogue partners to repeat previously stated information can be uncommon in natural dialogues. We aim to collect RGS-generated contradictory responses by identifying RGS response contradictions to follow-up questions. These follow-up questions are inquiries that seek additional information related to the information previously stated by the dialogue partner. These types of questions commonly arise during dialogues. They are similar to queries requesting repetitions of previously mentioned facts or opinions because they both seek information related to the previously stated content.

4.3 Dataset construction

In this study, we will showcase the vital importance of employing extensive datasets containing RGS-generated contradictory and noncontradictory responses to improve contradiction detectors. As stated earlier, large-scale data are currently lacking. To address this issue, we first performed an extensive collection of RGS-generated instances. This section outlines the methodology employed to build our dataset, followed by the detailed settings and the data collection outcomes for this study.

4.3.1 Construction method

Figure 4.1 illustrates our data collection process. We first prepare dialogue contexts as the inputs and then collect their RGS responses. The collected responses are classified into two groups, contradictory or noncontradictory, according to context. This process is based on that used by Nie et al. (2021), with the only differences being the approach to the dialogue

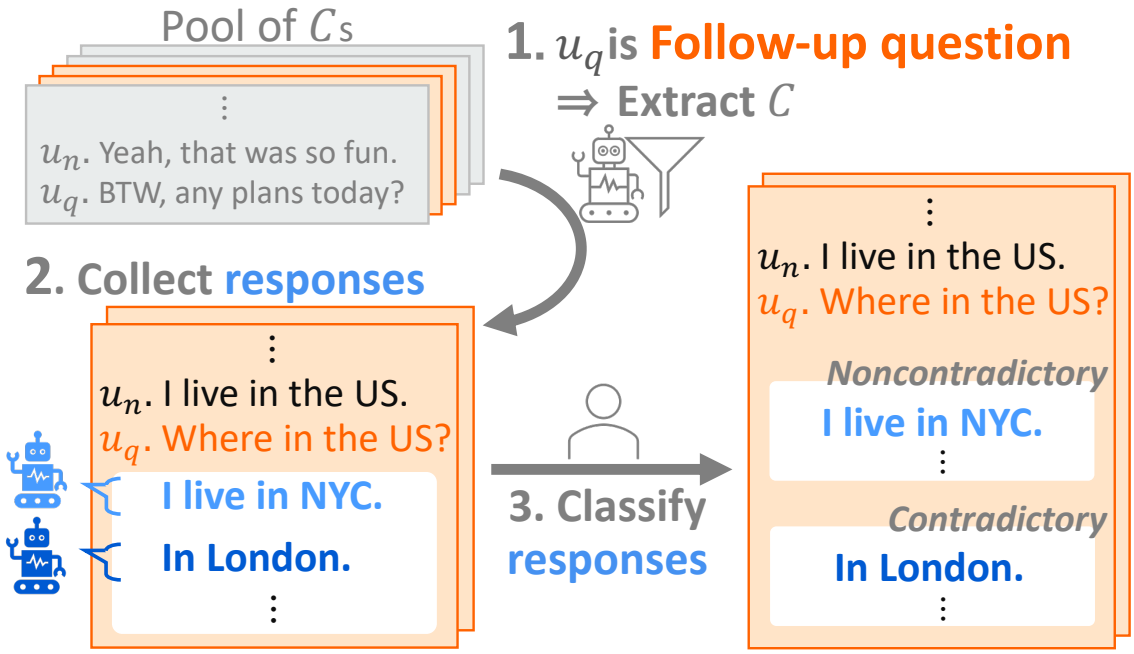

Figure 4.1 Overview of our data collection process.

context preparation and the focus on RGS-generated responses instead of human-written ones.

Method of dialogue context preparation

Contradictory responses are inconsistent with dialogue contexts; hence, their appearances depend on their contexts. For instance, it is improbable that a contradictory response will be generated in a dialogue context, in which only greetings are exchanged. We gathered follow-up questions (FQs) as the prime contexts for eliciting contradictions based on previous insights (Section 4.2.2) and our own small-scale analysis (Appendix 4.7.1).

RGSs do not generate contradictory responses exclusively to FQs; hence, addressing all contradiction types solely by collecting the contradictions to FQs is impractical. Nevertheless, initially concentrating on the representative inputs to efficiently gather a substantial number of contradictory responses is crucial. We believe that by refining contradiction suppression techniques using a large number of contradictory responses to these representative inputs, we can establish a groundwork for attempts to address contradictions in a broader input range.

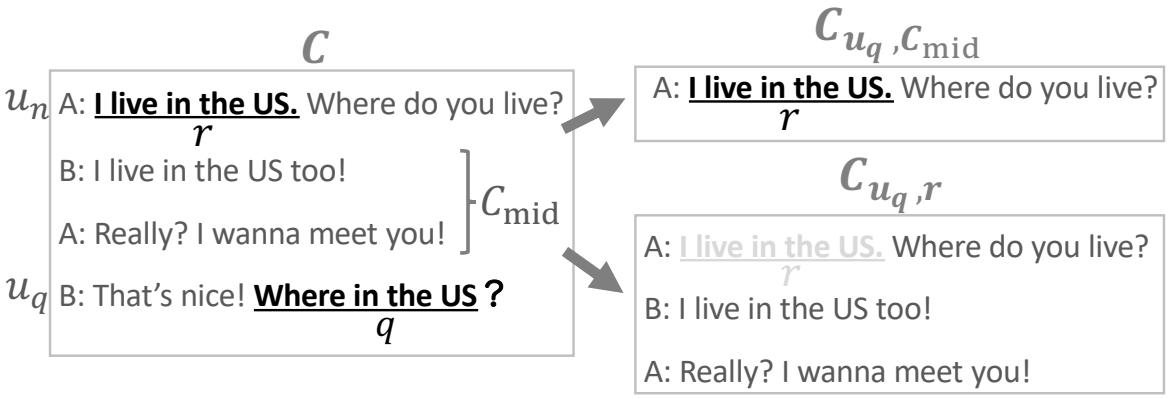

Figure 4.2 Example of $C_{u_q, C_{\text{mid}}}$ and $C_{u_q, r}$.

Idea for collecting FQ. In Figure 4.2, C refers to a dialogue context comprising more than n utterances[1] and concluding with an utterance u_q that contains a question q. We use $u_n \in C$ to represent the utterance that precedes u_q by n utterances. Note that we only consider scenarios wherein the u_n and u_q speakers are distinct individuals. When q is a question that refers to a specific segment r in u_n,[2] as illustrated on the left side of Figure 4.2, we regard q as an FQ for r. Throughout this paper, the segment r to which q refers will be termed "the referent of q." To determine whether r is the referent of q, i.e., whether q is an FQ for r, we must check if there is relevance between r and q. Fortunately, recent neural RGSs can generate highly relevant responses to contexts (Adiwardana et al., 2020; Zhang et al., 2020); hence, these RGSs are expected to capture the relevance between utterances effectively. In other words, if a large-scale neural RGS deduces the strong relevance between r and q, we can reasonably consider q as an FQ for r. We introduce herein a new automatic metric that employs a neural RGS[3] to assess the relevance between q and r.

Method for collecting FQ. As Figure 4.2 shows, $C_{u_q, C_{\text{mid}}}$ refer to C, excluding u_q, and the intervening utterances C_{mid} between u_n and u_q. Similarly, $C_{u_q, r}$ represents C with both u_q and r removed. If q is an FQ for r, it is improbable for $P(u_q | C_{u_q, C_{\text{mid}}})$ to exhibit a decrease compared to u_q's original conditional probability. Moreover, $P(u_q | C_{u_q, r})$ is likely lower than u_q's original conditional probability. Consequently, the following value is deemed high when q is an FQ for r:

$$\text{FQness} = P(u_q | C_{u_q, C_{\text{mid}}}) / P(u_q | C_{u_q, r}).$$

[1] Note that this study's term "utterance" refers to all sentences within a single turn.
[2] Consider q as question sentences in u_q and r as non-question sentences in u_n.
[3] This work employed Blenderbot (Roller et al., 2021), a well-known high-performance RGS.

Table 4.2 Number of contradictory responses to Cs extracted by RANDOM and TOP. Each value denotes the count of responses judged contradictory to u_ns by at least T workers out of 10.

(a) $n = 1$

	$T = 1$	$T = 2$	$T = 3$	$T = 4$
RANDOM	194 / 700	67 / 700	33 / 700	11 / 700
TOP	**238** / 700	**101** / 700	**50** / 700	**23** / 700

(b) $n = 3$

	$T = 1$	$T = 2$	$T = 3$	$T = 4$
RANDOM	246 / 700	77 / 700	31 / 700	10 / 700
TOP	**270** / 700	**141** / 700	**81** / 700	**43** / 700

(c) $n = 5$

	$T = 1$	$T = 2$	$T = 3$	$T = 4$
RANDOM	98 / 350	20 / 350	6 / 350	3 / 350
TOP	**126** / 350	**50** / 350	**25** / 350	**17** / 350

Here, we compute the probabilities using an RGS. We collect the FQs by selecting samples with the highest FQness from a pool of Cs.

Preliminary experiment for validating the effectiveness of FQness. We initially extracted samples from a pool of Cs for three different cases (i.e., $n = 1$, $n = 3$, and $n = 5$) by random sampling (RANDOM) and choosing samples with the highest FQness (TOP). Subsequently, we employed seven RGSs to generate responses to the Cs collected by RANDOM and TOP. We then compared the number of the RGS responses contradictory to the Cs obtained through the two abovementioned methods. Appendix 4.7.2 presents the detailed settings. Table 4.2 displays the comparison results, which confirmed that more contradiction labels were assigned to the C responses with a high FQness. This observation underscored the tendency of Cs with a higher FQness to provoke more RGS contradictions.

Method for RGS response collection

For every gathered C with a high FQness, multiple RGSs are utilized to collect the responses to gather diverse contradictory responses from various RGSs efficiently.

Method for RGS response annotation

We assign three human workers to assess each generated response and categorize it into two groups, contradictory and noncontradictory, according to their preceding referent r in u_n. If at least two workers determine the presence of contradictions in a response, this response is labeled as contradictory. If all workers agree that a response is consistent, this response is labeled as noncontradictory.

4.3.2 Construction settings

Settings of dialogue context preparation

A pool of Cs is formed by extracting n or more consecutive utterances from a dialogue corpus, ensuring that the final utterance includes questions. From this pool, we gather those with the highest FQness scores. For this study, we gathered Cs from the Multi-session Chat (MSC) dataset (Xu et al., 2022). The MSC dataset possesses the following characteristics, making it an ideal source for collecting Cs:

- low noise (e.g., few misspellings), and
- realistic dialogues between acquaintances, wherein speakers engage in in-depth discussions on a wide range of topics.

Based on the investigation described in Section 4.3.1, contradictions with $n = 5$ were less common than those with $n = 1, 3$, as evidenced by the `RANDOM` counts in Table 4.2. Furthermore, the annotation cost increased as the n value increased. Consequently, for this study, the highest value assigned to n was 5, and we intensively gathered FQs with $n = 1, 3$ to prepare for collecting contradictory responses. From the set of approximately 59k Cs in the MSC dataset, we extracted 3250, 1000, and 100 samples for $n = 1$, $n = 3$, and $n = 5$, respectively, based on the FQness scores.

Settings for RGS response collection

When gathering the responses for the extracted Cs, we employed a set of eight recent high-performance neural RGSs: Plato-2 (P2) (Bao et al., 2021), Plato-XL (PX) (Bao et al., 2022), Blender1-3B (B1) (Roller et al., 2021), Blender2-3B (B2) (Komeili et al., 2022; Xu et al., 2022), Blender3-3B (B3) (Shuster et al., 2022), Blender3-30B (BL), Opt-66B (O6) (Zhang et al., 2022), and ChatGPT (CG).[4] Each RGS generated one response to an input, resulting in eight responses for C. Appendix 4.7.3 presents the detailed settings.

[4]https://openai.com/chatgpt.

Table 4.3 Summary of our dataset. "# of C" and "# of N" denote the numbers of the collected contradictory and noncontradictory responses, respectively. The values in parentheses refer to the number of unique contexts.

Val. of n	# of C	# of N
1	8108 (2703)	12471 (2920)
3	2175 (739)	4378 (953)
5	220 (74)	422 (94)
Total	10503 (3516)	17271 (3967)

Settings for RGS response annotation

We employed Amazon Mechanical Turk (AMT)[5] for the worker recruitment. We ensured the creation of a cost-effective and high-quality dataset by carefully selecting highly skilled workers. We first presented a task with obviously correct answers. It contained 21 dialogue responses requiring classification into contradictory or noncontradictory according to their preceding referent r. We exclusively handpicked workers who scored fewer than two incorrect answers in this task. During the data collection phase, we published human intelligence tasks (HITs) to the selected workers, which required classifying 40 responses for five Cs.

4.3.3 Construction results

Table 4.3 reports several dataset statistics. Throughout the annotation process, the groups of three workers achieved average Fleiss' kappa values of 0.405, 0.465, and 0.408 for $n = 1$, $n = 3$, and $n = 5$, respectively. Given the intricacies involved in identifying contradictions, the substantial level of consensus signified the successful creation of a high-quality dataset. Table 4.1 provides examples of our dataset. Each sample comprises a dialogue context C containing u_n and u_q and a RGS response contradictory or noncontradictory to r in u_n with the labels assigned by the three workers.

4.4 Experiments

This section presents compelling evidence to support the hypothesis that employing the RGS-generated contradiction collection as a training resource yields notable enhancements in the performance of data-driven contradiction detectors.

[5] www.mturk.com.

Existing detectors have been developed by employing automatically synthesized or human-written contradictions as substituting training resources for RGS contradictions. Our experiments validate the potency of our dataset by assessing the contradiction detection performance of a detector trained on our dataset against that of a detector trained with human-written contradictions.

4.4.1 Experimental settings

Contradiction detector model. We conducted a performance analysis of a detector that underwent training on our dataset, juxtaposed with a detector fashioned similarly to the state-of-the-art model devised by Nie et al. (2021). Their detector was developed by fine-tuning RoBERTa (Liu et al., 2019) on the DECODE dataset, a collection of human-written contradictory and noncontradictory responses compiled by Nie et al. (2021). It was trained specifically for binary classification tasks requiring the prediction of consistency within a pair of given utterances. Following their settings, we developed a Contradiction Detector by fine-tuning RoBERTa on our dataset, denoted as $\mathrm{CD}_{\mathrm{RGS}}$. Similarly, we constructed a rival detector, $\mathrm{CD}_{\mathrm{HUM}}$, using an equivalent number of instances from the DECODE dataset as $\mathrm{CD}_{\mathrm{RGS}}$.

Training data for $\mathrm{CD}_{\mathrm{RGS}}$. Our dataset contains both contradictory and noncontradictory responses from eight RGSs. Our experiment performed a cross-validation test by selecting one RGS (i.e., target RGS) and using its responses as the test data. The samples excluding the target RGS's responses were used for training. We realized a comprehensive assessment of the detectors' performance by conducting the evaluation process eight times, varying the target RGSs each time. When we selected B2 as the target model, the number of training data samples was minimized to 8023 contradictory and 8023 noncontradictory responses; we reduced the number of training data samples to align with this number when we specified one of the other RGSs as a target RGS. Appendix 4.7.4 presents the training details.

Training data for $\mathrm{CD}_{\mathrm{HUM}}$. We randomly selected 8023 contradictory and 8023 noncontradictory human-written responses from the DECODE dataset. Other settings are the same as for $\mathrm{CD}_{\mathrm{RGS}}$.

In-domain test sets. We randomly selected 100 contradictory and 100 noncontradictory responses of the target RGS responses from our dataset as test samples. Note that a training set might also contain responses from other RGSs that share the same contexts as these 200 test samples. We excluded these corresponding samples from the training set to ensure a

fair evaluation of the detectors' ability to identify contradictions from unknown RGSs in unfamiliar contexts.

Out-of-domain test sets. The above RGS-generated test sets were derived from the corpus used to develop the training set for $\mathrm{CD}_{\mathrm{RGS}}$. Furthermore, these sets exclusively comprise responses to FQs. In order to assess the detector's effectiveness in identifying contradictions in RGS responses to non-FQ contexts from unfamiliar dialogue corpora, we prepared two out-of-domain test sets. One set originated from the Topical-Chat dataset (Gopalakrishnan et al., 2019), and the other from the DailyDialog dataset (Li et al., 2017). Each of these sets comprises seven subsets, each containing 50 contradictory and 50 noncontradictory responses from P2, PX, B1, B2, B3, BL, or O6.[6] The contexts in these sets were randomly selected from all contexts in the corpora concluding with an utterance containing questions not limited to FQs.[7] Appendix 4.7.5 details the construction process. The utilization of the subsets of these two test sets was the same as that of the subsets of the in-domain test set, except that even the subsets comprising non-target RGSs' responses were excluded from the training set to prevent detectors from being trained on the same domain data. In addition, we employed Nie et al. (2021)'s Human-Bot dataset, which possesses 382 contradictory and 382 noncontradictory RGS responses in human-bot dialogues.

Human-written test set. We utilized Nie et al. (2021)'s By-Human test set comprising 2108 contradictory and 2108 noncontradictory human-written responses. This allowed us to verify that $\mathrm{CD}_{\mathrm{HUM}}$ was reasonably well-trained in our settings, although addressing human-written contradictions falls beyond the scope of our study.

4.4.2 Results

Table 4.4 (a), (b), and (c) display the accuracy of the contradiction detectors in solving whether a given response contradicts the specific preceding utterance u_n for the human-written, in-domain, and out-of-domain test sets, respectively.

(a) Human-written test set. $\mathrm{CD}_{\mathrm{HUM}}$ obtained a high accuracy of 0.952 on the By-Human test set, confirming that $\mathrm{CD}_{\mathrm{HUM}}$ was properly trained.

[6] CG was omitted from the test set construction due to cost considerations, as the frequency of contradictions from CG was relatively low.

[7] Considering that non-question contexts may allow contextually irrelevant replies, such as prompting changes in the topic, we anticipate a lower occurrence of contradictions. Our emphasis on responses only to questions aligns with cost considerations.

Table 4.4 Accuracy of the detectors for (a) human-written, (b) in-domain, and (c) out-of-domain test sets. $\mathrm{CD}_{\mathrm{RGS}}$'s score for By-Human is the median of $\{0.819, 0.827, 0.838, 0.843, 0.847, 0.857, 0.859, 0.871\}$ since we trained the eight $\mathrm{CD}_{\mathrm{RGS}}$ detectors. $\mathrm{CD}_{\mathrm{RGS}}$'s score for Human-Bot refers to the accuracy of the one trained without B1's responses among the eight $\mathrm{CD}_{\mathrm{RGS}}$ detectors because Human-Bot contains B1's responses.

(a) Human-written test set.

Detector	By-Human
$\mathrm{CD}_{\mathrm{HUM}}$	**.952**
$\mathrm{CD}_{\mathrm{RGS}}$	.845

(b) In-domain test sets. Scores for each target RGS are presented.

Detector	P2	PX	B1	B2	B3	BL	O6	CG
$\mathrm{CD}_{\mathrm{HUM}}$	.600	.575	.615	.540	.655	.555	.565	.650
$\mathrm{CD}_{\mathrm{RGS}}$	**.800**	**.725**	**.715**	**.750**	**.765**	**.745**	**.690**	**.790**

(c) Out-of-domain test sets. For the Topical-Chat and DailyDialog test sets, scores for each target RGS are presented.

	Test set from Nie+'21	Test sets from Topical-Chat / DailyDialog						
Detector	Human-Bot	P2	PX	B1	B2	B3	BL	O6
$\mathrm{CD}_{\mathrm{HUM}}$	.749	.55/.52	.58/.60	.61/.55	.60/.59	.68/.61	.67/.55	.59/.53
$\mathrm{CD}_{\mathrm{RGS}}$	**.787**	**.77/.77**	**.72/.67**	**.74/.68**	**.70/.72**	**.73/.76**	**.82/.64**	**.81/.75**

(b) In-domain test sets. However, $\mathrm{CD}_{\mathrm{HUM}}$ achieved low accuracy for the subsets of our RGS-generated dataset. Particularly, it had an accuracy of only 0.540 when B2 was the target RGS, which is problematic in practical applications. In contrast, $\mathrm{CD}_{\mathrm{RGS}}$ gained higher accuracy on our RGS-generated test sets. The training process for $\mathrm{CD}_{\mathrm{RGS}}$ excluded any contradiction data from the target RGSs and samples that shared the same dialogue contexts as the test data, thereby effectively detecting contradictions from unknown RGSs when confronted with unfamiliar contexts.

(c) Out-of-domain test sets. For all three test sets, the performance of $\mathrm{CD}_{\mathrm{RGS}}$ significantly outperformed $\mathrm{CD}_{\mathrm{HUM}}$. These results emphasize that detectors trained on our dataset can effectively detect contradictions in RGS responses to out-of-domain and non-FQ contexts.

Note that it has been confirmed that $\mathrm{CD}_{\mathrm{RGS}}$ exhibited superior performance even when the entirety of DECODE's samples was employed for training $\mathrm{CD}_{\mathrm{HUM}}$ (Appendix 4.7.6), although the above experiments employed only approximately half of DECODE's samples during $\mathrm{CD}_{\mathrm{HUM}}$'s training.

Table 4.5 Example of Opt-66B's contradictory responses with an intra-utterance inconsistency.

Context
A: I made plans to travel to a new place next month. B: What attracted you to this new place? Where is it?
RGS responses on speaker A's side
O6: I've been to this place before and I really liked it. It's in a country I've never been to before.

4.5 Analysis

In the previous section, we confirmed that using the contradictory responses generated by the RGS as training data improves the accuracy of the contradiction detector. However, the accuracy still varies from 100%, leaving room for improvement. To further improve performance, it may be necessary not only to collect diverse and large amounts of contradictory responses by RGSs but also to analyze the characteristics of these responses and build a data-driven method better suited for identifying them. Conveniently, our collections can also be used for this type of analysis. In this section, we describe the results of our analysis of the characteristics of the contradictory response by RGS using our collected collections, which may be used to devise innovative systems to identify RGS-generated contradictions. Section 4.5.1 delves into the intrinsic features of the generated responses, while Section 4.5.2 presents an examination centered around the dialogue contexts that trigger RGS contradictions.

4.5.1 Analysis of RGS-generated responses

Our analysis revealed two types of characteristic contradictions in the RGS-generated responses: contradictions arising from intra-utterance inconsistencies and those involving ambiguous expressions.

Contradiction arising from intra-utterance inconsistency

A qualitative analysis of the RGS-generated contradictory responses in our dataset revealed that one distinctive way contradictions occur is through intra-utterance inconsistency. Table 4.5 provides an example of this contradiction type. In this conversation, Opt-66B generated inconsistent information within a single utterance, saying, "I've been to this place" while also stating, "It's in a country I've never been to." When conflicting information co-

Table 4.6 Example of Plato-2's contradictory responses with ambiguity. The determination of whether or not a contradiction exists hinges upon the interpretation assigned to the bolded term "interview," particularly if it is construed to differ from the preceding interview.

Context
A: I had a promising interview today! B: Oh excellent! How did it go, what made it so excellent?
RGS responses on speaker A's side
P2: i think i did well because they called me back to set up **an interview**.

exists within a statement, at least one of them contradicting the context becomes highly probable. Instances of contradictions stemming from intra-utterance inconsistencies are occasionally observed across multiple RGSs. To delve deeper into this phenomenon, we counted the inconsistencies among the 50 randomly selected contradictory responses in our dataset for each of the eight RGSs. Our findings indicated that seven RGSs generated at least one contradictory response featuring an intra-utterance inconsistency. Conversely, none of the 50 randomly sampled human-written contradictory responses in DECODE exhibited an intra-utterance inconsistency. These results suggest contradictory responses featuring intra-utterance inconsistencies are particularly frequent in RGS responses.

Contradiction involving ambiguous expression

We observed a notable distinction in the human annotation tendency on the existence of contradictions between the set of human-written responses in DECODE and our compilation of RGS-generated responses. Both our study and that of Nie et al. (2021) employed a similar approach in selecting the human workers who identified the contradictory responses during the data creation process (Section 4.3.2). However, within the subset of instances where at least one worker detected the contradictions, a significant gap was observed in the proportions where the other two workers also concurred on the existence of contradictions. This proportion was 78.4% for the human-written responses and 30.4% for the RGS-generated ones. This dissimilarity could have stemmed from the RGS's propensity to generate ambiguous expressions concerning consistency, as demonstrated in Table 4.6. Such responses appeared to result in differing judgments regarding the presence of contradictions, depending on how individual workers interpreted them. Addressing these contradictory responses is crucial, even if some workers may miss the inconsistencies, because these responses, once perceived as contradictory by actual users, can significantly detriment the dialogue quality.

Effects on data-driven approaches

The experiments in Section 4.4 exhibited that the detector training with the RGS-generated contradictions led to a noticeable enhancement in the detectors' capability to identify RGS contradictions, although it was far from the ideal accuracy. We hypothesized that this outcome could be attributed, at least in part, to the training on the RGS-generated instances, which facilitated the acquisition of identifying features typical of RGS-generated contradictions, encompassing those expounded upon in Sections 4.5.1. If this hypothesis is correct, then identifying the features typical of RGS-generated contradictory responses and then concentrating the collection of training data on responses with these characteristics may be important for achieving a more accurate contradiction detector. We investigated the hypothesis' validity by taking one of the two abovementioned contradiction types, contradiction with ambiguous expression, as examples.

Analysis method. Our experiments revealed that 1377 RGS contradictory responses from our in-domain test sets and the validation sets used when training $\mathrm{CD}_{\mathrm{RGS}}$ (Appendix 4.7.4) were missed by $\mathrm{CD}_{\mathrm{HUM}}$ but successfully flagged by $\mathrm{CD}_{\mathrm{RGS}}$. Plausibly, some of these instances may exhibit certain features inherent to the RGS contradictions, which the training with RGS-generated data facilitated $\mathrm{CD}_{\mathrm{RGS}}$ to recognize. Therefore, we investigated if they encompass the distinguishing characteristic, i.e., ambiguous expression.

Results. All contradictory samples in our dataset were deemed contradictory by either two or three workers. The instances judged contradictory by only two workers may encompass ambiguities regarding consistency. Within those above 1377 contradictory responses, the proportion of the samples classified as contradictory by only two workers amounted to 51.1%. Conversely, among the 4394 contradictory responses from our validation and test sets that both $\mathrm{CD}_{\mathrm{HUM}}$ and $\mathrm{CD}_{\mathrm{RGS}}$ successfully identified, only 43.4% of the samples were determined contradictory by two workers. This proportion gap exhibited statistical significance at the 1% significance level in the chi-square test, underscoring that training on RGS-generated data enhanced the detector's capacity to recognize the contradictions characterized by ambiguity.

4.5.2 Analysis of dialogue contexts

If specific dialogue contexts induce contradictory responses from various RGSs, identifying their contributing characteristics may become crucial in developing more effective techniques for identifying contradictions. Our dataset is suitable for this investigation because

Table 4.7 Example of Plato-2's contradictory responses containing a partner's bolded statement.

Context
A: Have you taken any new pictures?
B: I managed to get out at the weekend and get loads of shots in the snow we had. [. . .]
A: Oh wow you had snow!? **We just had rain all weekend :)** [. . .] Did you have a nice chilled weekend? [. . .]
RGS responses on speaker B's side
P2: it was a good weekend here, **we got to enjoy the cold rain**!

it contains a large number of Cs for which diverse RGSs generate responses. This section presents our findings, providing a quantitative validation for the presence of these contexts. We also discerned several attributes associated with them.

Existence of contexts from which multiple RGSs generate contradictions

For every RGS employed in our data collection, we classified each dialogue context within our dataset based on whether or not it induced contradictions from that particular RGS. We then assessed the degree of agreement on the classification of all RGSs utilizing Fleiss' kappa, obtaining a result of 0.098 that demonstrated a slight degree of accord in line with the criteria by Landis and Koch (1977). This outcome signified the presence of specific dialogue contexts from which multiple RGSs generated contradictions.

Features of these dialogue contexts

A further examination showcased certain C features in which a relatively large number of RGSs generated contradictions in our data collection.[8] Our analysis was centered around dialogue act labels and lexical attributes, which are highly interpretable features and appear well-suited as the first analysis's focal points.

Analysis results of dialogue acts. When we assigned dialogue act labels to u_qs in our dataset,[9] we observed a notable trend, that is, u_qs categorized as "Declarative Yes-No-

[8]When classifying Cs into two sets according to the presence of a feature, a statistically significant disparity in the average count of the induced contradictory responses from the eight RGSs per C between the two sets was unveiled through a one-tailed t-test at 1% significance level.

[9]We trained RoBERTa on the Switchboard corpus (Jurafsky et al., 1997) to develop a dialogue act labeler. The test set accuracy was 80.4%.

Questions" or "Statement-non-opinion" were more prone to triggering contradictory responses. For the former label, among the 193 assigned instances, the average count of the contradictory responses from the eight RGSs per C was 2.77. The average for the 4084 unassigned instances was lower at 2.41. This phenomenon could have arisen from a deficiency in the RGS's ability to generate appropriate responses while being cognizant that a repetition of previous information is being solicited. For the latter label, focusing on 2118 assigned instances indicated a higher average of 2.49 contradictory responses compared to an average of 2.36 for the 2159 unassigned ones. This disparity could arise from the RGSs' inability to differentiate between the content of the dialogue partners' statements and their own utterances. Hence, RGSs may generate responses incorporating the partners' information as if it were their own, even if it is inconsistent with their past statements, as exemplified in Table 4.7.

Analysis results of lexical features. The u_qs containing the interrogative term "how" can provoke contradictions. More precisely, the mean count of the contradictory responses within the 764 applicable contexts stood at 2.60, while that in the 3513 inapplicable contexts was 2.39. Expounding upon methods or the extent regarding a subject while upholding consistency within the context poses a challenge for the current RGSs.

4.6 Conclusion

No attempt has been made to build an extensive collection of RGS-generated contradictory responses, which results in the scarcity of training data for contradiction detectors.

In this study, we initially built an extensive collection of contradictory and noncontradictory responses generated by various high-performance RGSs. We then showed that a contradiction detector trained on human-written contradictions exhibits low accuracies when detecting actual RGS contradictions and validated that training on our collection improves this situation. We performed the contradiction detector training to demonstrate the effectiveness of leveraging our dataset to improve data-driven contradiction suppression methods. However, our gathered contradictory and noncontradictory responses can also be employed in other data-driven approaches. We also comprehensively analyzed our collection from various angles, producing valuable insights into the RGS-generated contradictory responses, which we believe are crucial for effective contradiction identification.

Future challenges include collecting data with a broader context variety than the follow-up questions.

4.7 Appendix

4.7.1 FQ analysis in existing dataset

We randomly examined 50 responses from a pool of 382 RGS-generated contradictory responses in the human-bot dialogues collected by Nie et al. (2021). Remarkably, 25 of these 50 contradictory responses were elicited by the FQs, strongly indicating that the FQ plays a prominent role in provoking RGS contradictions.

4.7.2 Preliminary experiment settings

Settings of dialogue context preparation. We utilized the pool of Cs described in Section 4.3.2 as the source for extracting instances using `RANDOM` and `TOP`. Each `TOP` and `RANDOM` extracted 100, 100, and 50 samples from the pool for $n = 1$, $n = 3$, and $n = 5$, respectively.

Settings for RGS response collection. Seven RGSs were employed to generate the responses for each C: Plato-2, Plato-XL, Blender1-3B, Blender2-3B, Blender3-3B, Blender3-30B, and Opt-66B.[10] We specifically had each RGS generate 100 response candidates for each input through top-p sampling, with a value of p set to 0.5. We chose the response with the highest generation probability among the 100 candidates.

Settings for RGS response annotation. Each RGS response was manually assessed to determine its consistency with the context. The 10 AMT workers assigned to each response performed a binary classification task to distinguish between the contradictory and noncontradictory responses. We solely focused on evaluating the consistency with u_n due to cost considerations.

4.7.3 Settings for dataset construction

Each of the eight RGSs generated one response to an input, resulting in eight responses for each C. We enhanced the efficiency of gathering the contradictions by choosing the final response of a RGS to an input from the top 100 candidates with the highest contradiction probability predicted by the state-of-the-art contradiction detector (Nie et al., 2021). We utilized top-p sampling (Holtzman et al., 2020) to collect the 100 candidates. We set a value of p to 0.5, which was lower than the default value of 0.9 used in major platforms, such as ParlAI (Miller et al., 2017), to avoid sampling responses with low generation probabilities.

[10]Note that the ChatGPT API service was not yet available when the preliminary experiment was conducted.

This allowed us to gather candidates with a high generation probability by the RGS and a high likelihood of being contradictory. We employed OpenAI's API[11] for ChatGPT, Knover[12] for Plato-2 and Plato-XL, and ParlAI for the others.

4.7.4 Settings for detector training

Samples for training. A negative pair comprised a contradictory response in our dataset and the preceding utterance u_n. In contrast, a positive pair comprised a noncontradictory response from our dataset and one randomly selected from its preceding utterances by the same speaker. This was because the responses annotated as noncontradictory with u_ns were also likely to be noncontradictory with the other preceding statements. By introducing randomness into the selection of preceding utterances for pairing with a noncontradictory RGS response, we aimed to create positive pairs comprising unrelated utterances. These pairs could be valuable for training detectors to recognize that unrelated pairs should be categorized as noncontradictory.

Hyperparameters. We trained detectors by employing the implementation of Hugging Face (Wolf et al., 2020) with its default settings, excluding a few parameters.[13] We updated the model parameters until we reached a point where early stopping was triggered. Early stopping was determined by assessing the validation data accuracy, a distinct subset comprising 10% of the training data and withheld from the training process. We saved the model with the highest accuracy on the validation data at each learning rate and ultimately selected that with the highest validation accuracy among all the saved models.

4.7.5 Settings for test set construction

We first constructed two pools of C by extracting $n = 1, 3$ or more consecutive utterances from the Topical-Chat dataset and the DailyDialog dataset, respectively, ensuring that the final utterance contains questions. From each of the two pools, we randomly sampled those consecutive utterances. Specifically, for the Topical-Chat dataset, we sampled 300 and 100 samples from the pool for $n = 1$ and 3, respectively. Similarly, for the DailyDialog dataset, we sampled 200 and 100 samples from the pool for $n = 1$ and 3, respectively. The other settings are the same as in our large-scale dataset construction described in Section 4.3, ex-

[11] https://platform.openai.com.

[12] www.github.com/PaddlePaddle/Knover.

[13] `train_batch_size`: 128, `weight_decay`: 0.01, `eval_steps`: 200, `early_stopping_patience`: 1, and `learning_rate`: {1e-6, 5e-6, 1e-5, 5e-5}.

Table 4.8 Accuracy of $\mathrm{CD}_{\mathrm{RGM}}$ and $\mathrm{CD}_{\mathrm{HUMF}}$ for (a) human-written, (b) in-domain, and (c) out-of-domain test sets.

(a) Human-written test set.

Detector	By-Human
$\mathrm{CD}_{\mathrm{HUMF}}$	**.958**
$\mathrm{CD}_{\mathrm{RGM}}$	.845

(b) In-domain test sets. Scores for each target RGM are presented.

Detector	P2	PX	B1	B2	B3	BL	O6	CG
$\mathrm{CD}_{\mathrm{HUMF}}$	.625	.620	.565	.585	.595	.625	.575	.715
$\mathrm{CD}_{\mathrm{RGM}}$	**.800**	**.725**	**.715**	**.750**	**.765**	**.745**	**.690**	**.790**

(c) Out-of-domain test sets. For the Topical-Chat and DailyDialog test sets, scores for each target RGM are presented.

	Test set from Nie+'21	Test sets from Topical-Chat / DailyDialog						
Detector	Human-Bot	P2	PX	B1	B2	B3	BL	O6
$\mathrm{CD}_{\mathrm{HUMF}}$	**.829**	.57/.52	.64/.59	.67/.56	.64/.59	.69/.66	.71/.54	.59/.54
$\mathrm{CD}_{\mathrm{RGM}}$	.787	**.77/.77**	**.72/.67**	**.74/.68**	**.70/.72**	**.73/.76**	**.82/.64**	**.81/.75**

cept for the method of collecting C described above and that responses of ChatGPT were not collected.

4.7.6 Experiments with more human-written contradictions

Table 4.8 presents the outcomes of evaluating the contradiction detection capabilities of $\mathrm{CD}_{\mathrm{RGM}}$ in comparison to $\mathrm{CD}_{\mathrm{HUMF}}$. $\mathrm{CD}_{\mathrm{HUMF}}$ was trained using all human-written data from the DECODE dataset, consisting of 15605 contradictory and 15605 noncontradictory responses, following the same training procedure as $\mathrm{CD}_{\mathrm{HUM}}$. The comparison methodology aligns with the experimental approach outlined in Section 4.4.

Noteworthy is the observation that, despite $\mathrm{CD}_{\mathrm{RGM}}$'s training dataset being approximately half the size of $\mathrm{CD}_{\mathrm{HUMF}}$, it demonstrated superior performance across all RGM-generated test sets, with the exception of the Human-Bot set. The Human-Bot test data comprises only a limited number of conversational exchanges during initial encounters. Given that $\mathrm{CD}_{\mathrm{HUMF}}$'s training dataset also encompasses human-written contradictory responses following first-meeting dialogues, it is conceivable that the overlap in domains enabled the detector to recognize contradictions in the Human-Bot test data.

Chapter 5

Expanding evaluation target to n-best responses

5.1 Introduction

An RGS finally outputs one response to a given context, but this does not necessarily mean that the RGS assumes only one response candidate for the final output. Rather, in recent years, RGS usually outputs a final response by selecting the candidate from an n-best candidate list (Nie et al., 2021; Welleck et al., 2019). Prior work has demonstrated that generating the n-best lists with noncontradictory 1-bests is an open challenge (Kim et al., 2020; Li et al., 2021; Nie et al., 2021). Thus, one practical technique for avoiding contradiction is to have an accurate contradiction detector that eliminates all contradictory candidates from the n-best list (Nie et al., 2021). In this scenario, the consistency of all candidates in the n-best list, not just the 1-best, substantially impacts whether the final output is contradictory because the final response is chosen from the n-best list. Therefore, the evaluation of CA in anticipation of the practical use of RGS must take into account not only the 1-best but also all the n-best candidates. Nonetheless, earlier quantitative investigations of contradiction relied solely on 1-bests from RGSs (Li et al., 2021).

In this study, we propose a framework to evaluate RGSs' more practical CA by taking the consistency of all of the n-best candidates into account, assuming the post-processing where an ideal contradiction detector chooses noncontradictory candidates from the n-best candidates. Specifically, we first consider how evaluating an n-best list should be approached. Then, we propose a method for statistically evaluating the n-best lists (Figure 5.1). Since it is impractical to study all conceivable contradictions in a dialogue, we first focus on contra-

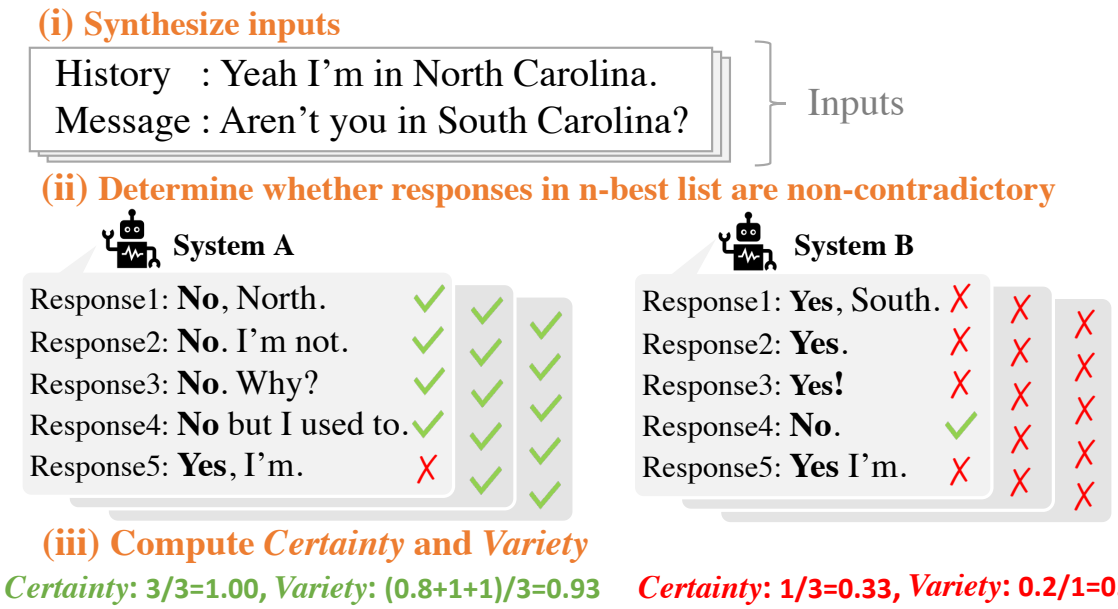

Figure 5.1 Overview of our evaluation framework. The framework evaluates n-best lists by (i) synthesizing a stimulus input that induces contradictions, (ii) automatically determining whether responses in the n-best lists are contradictory, and (iii) computing *Certainty* and *Variety*.

dictions in response to polar questions.[1] We use our method to highlight the CA of recent high-performance neural RGSs and methodologies. Our results show that beam search has limitations in terms of avoiding contradiction and that the newer techniques, such as unlikelihood training (Welleck et al., 2020), can help overcome these limitations.

5.2 Evaluation perspectives

5.2.1 Proposed metrics

First, n-best lists must be generated to prevent contradiction, assuming the filters can remove contradictory responses. An ideal RGS produces output that is noncontradictory and high quality in many other criteria, such as relevance or informativeness. An RGS must generate at least one noncontradictory candidate to deliver a noncontradictory output. Furthermore, even noncontradictory candidates could be eliminated based on other criteria (e.g., relevance, informativeness). Therefore, it can be hypothesized that having more noncontradictory responses in an n-best list would enhance the final output quality across various criteria. Tak-

[1] Codes and test set are available at https://github.com/shiki-sato/nbest-contradiction-analysis.

ing the above into account, we examine n-best lists based on the certainty of the existence of noncontradictory responses (***Certainty***), and the variety of noncontradictory responses (***Variety***):

- ***Certainty***: The proportion of the n-best lists that have at least one noncontradictory response.
- ***Variety***: The proportion of noncontradictory responses in each n-best list when only the n-best lists with at least one noncontradictory response are collected.

Given a set of inputs $\mathcal{Q}$, we calculate them as follows:

$$\textit{Certainty} = \frac{|\mathcal{Q}'|}{|\mathcal{Q}|}, \textit{Variety} = \frac{1}{|\mathcal{Q}'|} \sum_{q \in \mathcal{Q}'} \frac{\mathrm{cnt}(f(q))}{|f(q)|}$$

$$\mathcal{Q}' = \{\, q \mid \mathrm{cnt}(f(q)) > 0, q \in \mathcal{Q} \,\}$$

where $f(\cdot)$ is an n-best list generation function and $\mathrm{cnt}(\cdot)$ is a function that returns the number of noncontradictory responses from a given n-best list. For example, the *Certainty* of an RGS that generates n-best lists with a combination of noncontradictory and contradictory responses is high, but its *Variety* is low. However, an RGS that always generates n-best lists with only noncontradictory or contradictory responses has a high *Variety* but a low *Certainty*. We anticipate that n-best lists must include noncontradictory responses (*Certainty*$= 1.0$), with a high proportion (high *Variety*).

5.2.2 Relation with existing metrics

The assessment of n-best output has been a focal point in the realm of natural language processing. Specifically within the domain of information retrieval, several approaches have been suggested to evaluate the validity of the n documents retrieved for a given query as suitable search results. The relationship between existing metrics for information retrieval and our proposed metrics is delineated below.

Recall. This metric evaluates the percentage of retrieved documents among all documents that should be retrieved. In the evaluation of response consistency, employing recall is difficult since it is not possible to define "all documents that should be included."

Precision. This metric evaluates the percentage of documents that should be retrieved among all retrieved documents. Our evaluation metrics allow for a more detailed analysis of

Table 5.1 Acquiring dialogue context by transforming the Natural Language Inference (NLI) data.

NLI data			Dialogue context for our test	
Entailment	Premise: yeah i'm in North Carolina Hypothesis: I'm in North Carolina.	⟶	**EntQ**	History: Yeah I'm in North Carolina. Message: Are you in North Carolina?
Contradiction	Premise: yeah i'm in North Carolina Hypothesis: I'm in South Carolina.	⟶	**CntQ**	History: Yeah I'm in North Carolina. Message: Aren't you in South Carolina?

what precision assesses. For example, when the average of precisions is high, our metrics can distinguish whether some of the n-bests have extremely high precision or the majority of the n-bests have a certain degree of precision.

Rank-aware metrics. Rank-aware metrics, such as MRR (Craswell, 2009), evaluate n-bests considering the rank of each item in the n-bests. Since dialogue response generation assumes the reranking of n-best items based on other perspectives, our task does not need to consider their ranks.

5.3 Inputs and evaluation

To evaluate an RGS from the aforementioned viewpoints, we consider how to prepare the inputs and evaluate the generated responses in this section.

5.3.1 Inputs for highlighting contradictions

Polar echo question. An *echo question* (Noh, 1998) confirms or clarifies the context information by repeating the utterance of another speaker. It is commonly used when the speaker does not hear or understand what was said correctly, or when the speaker wishes to express incredulity. Based on Li et al. (2021)'s discovery, contradictions emerge mostly when speakers refer to earlier information communicated in dialogue; we use echo questions as stimulus input in our evaluation to elicit contradictory responses. We use polar-typed echo questions to make our evaluation more succinct and quantitative. Since polar questions allow for basically only two responses, *yes* or *no*, we can clearly determine whether the generated response is contradictory or not. Furthermore, by analyzing the produced responses as a yes/no binary classification issue, it allows for quantitative discussion of experimental outcomes based on the probability level.

Input preparation. We use the dataset from the natural language inference (NLI) task to effectively obtain the inputs described in the preceding paragraph. This dataset specifies the logical relationship (i.e., entailment, neutrality, or contradiction) between a premise and its associated hypothesis. We transform the NLI dataset into dialogue data using a set of basic rewriting rules.[2] Our test involves two types of inputs, which can be classified as follows:

- EntQ: generating a *confirmation* response.
- CntQ: generating a *refutation* response.

Table 5.1 displays the input samples and how they are transformed from the initial NLI data. Each input is made up of the following two utterances: the history and message. In our evaluation, the RGS generates responses to a given message, assuming the RGS has generated the history in the preceding turn.

5.3.2 Contradiction detection for output

To compute the *Certainty* and *Variety*, we must first determine whether each generated response in the n-bests compared to the inputs is contradictory. The simplest method for detecting the contradictions is to check whether the response begins with *yes* or *no*. However, in the event of an indirect expression (e.g., *Why not?*), this method cannot detect the contradictions. Therefore, we use an automated yes-no classifier to categorize the n-best responses to EntQ/CntQ. We train the classifier by fine-tuning RoBERTa (Liu et al., 2019) using the Circa dataset (Louis et al., 2020), which comprises pairs of polar questions and indirect responses, as well as annotations for the answer's interpretation, to categorize utterances as affirmations or refutations.[3]

5.4 Experiments

We demonstrate how our framework shows the properties of n-best lists, which could be quite influential in terms of avoiding contradiction. We demonstrate this by comparing the n-bests generated by conventional beam search (BS) versus recently proposed techniques.

5.4.1 Experimental settings

Inputs preparation. We used the Multi-Genre NLI Corpus (Williams et al., 2018) to obtain inputs, which is a large scale and is consistent in good quality NLI data. We created

[2]The details are described in Appendix 5.6.1.
[3]The details are described in Appendix 5.6.2.

Table 5.2 *Certainty* and *Variety* of 10-best lists using beam search with beam size $B = 10$.

	Certainty		*Variety*	
RGS	EntQ	CntQ	EntQ	CntQ
Blender 400M	0.806	0.747	0.780	0.775
Blender 1B	0.832	0.752	0.832	0.753
Blender 3B	0.856	0.768	0.824	0.737
DialoGPT 345M	0.938	0.917	0.750	0.669
DialoGPT 762M	0.883	0.918	0.671	0.713

2,000 EntQ/CntQ inputs by extracting 2,000 samples labeled with *entailment* or *contradiction*.[4]

RGSs. We used the following two recently developed high-performance RGSs: DialoGPT (Zhang et al., 2020) and Blender (Roller et al., 2021).[5]

5.4.2 Evaluation of *n*-best using beam search

Let B denote the beam size during generation. It has been empirically found that using beam search with $B = 10$ to generate a response yields excellent quality results and has a frequently used value (Roller et al., 2021; Zhang et al., 2020). Table 5.2 displays the *Certainty* and *Variety* of 10-best lists generated using these methods. Figure 5.2 also depicts the *Certainty* and *Variety* of n-best lists generated using different beam sizes.

***Certainty*.** Table 5.2 illustrates that in approximately 10% of CntQ-type inputs, even the highest scoring RGS generates 10-best lists full of contradictory responses. Even with a perfect response filter, the RGSs are unable to provide noncontradictory answers to these questions. It should be emphasized that the error rate is not low, given that the inputs are polar questions with highly restricted viable responses. Expanding the beam size can increase the number of n-best lists with at least one noncontradictory response. Indeed, increasing the beam size enhances the *Certainty* ((a) and (b) in Figure 5.2). By increasing B to 40, the *Certainty* of using DialoGPT 345M for both EntQ- and CntQ-type inputs achieve 1.0.

***Variety*.** With $B = 10$, all the RGSs' *Variety* are more than 0.5 (chance rate) (Table 5.2). Therefore, rather than being fully random, the RGSs generate n-best lists with a degree of

[4]We used the samples in the Telephone domain; this domain covers open-domain conversations.

[5]The details of the settings are described in Appendix 5.6.3.

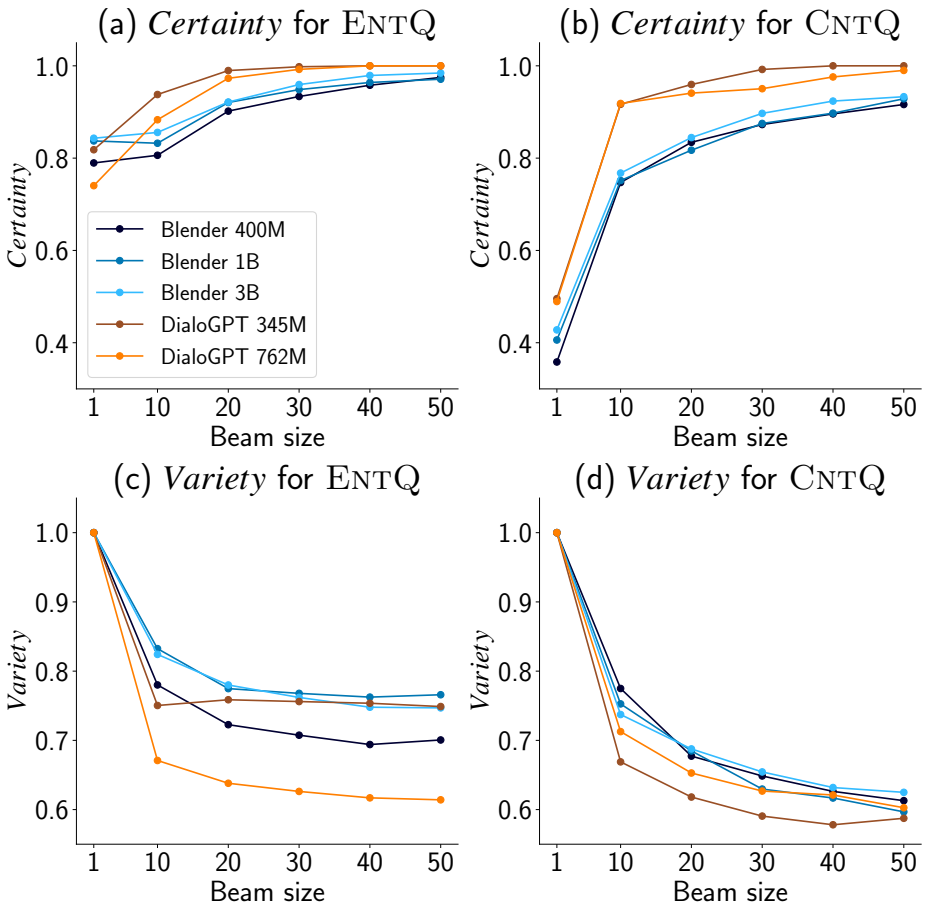

Figure 5.2 *Certainty* and *Variety* of n-best lists using beam search with various beam sizes.

directionality toward avoiding contradictions. However, increasing the size of beam reduces the *Variety* ((c) and (d) in Figure 5.2), resulting in lower output quality. For example, the *Variety* of DialoGPT 345M with $B = 40$ for CntQ-type inputs (an RGS with *Certainty* of 1.0 for both EntQ- and CntQ-type inputs) decreases to 0.58.

Overall. In terms of avoiding contradiction, our evaluation framework demonstrated the features of the n-best lists of the beam search. The *Certainty* did not achieve 1.0 in the commonly used configuration ($B = 10$). When the beam size is increased, the *Certainty* increases to 1.0, whereas the *Variety* reduces dramatically. These results show the trade-off between *Certainty* and *Variety* as a function of beam size; in this example, we found constraints in obtaining high *Certainty* and *Variety* with beam search. Furthermore, it is found that the *Certainty* obtained using DialoGPT is greater than that obtained using Blender, whereas the opposite is true for *Variety*, suggesting that various RGSs behave differently in

Table 5.3 *Certainty* and *Variety* of 10-best lists using various techniques with Blender 3B.

	Certainty		*Variety*	
Technique	EntQ	CntQ	EntQ	CntQ
BS	0.856	0.768	0.824	0.737
DBS	0.999	0.981	0.758	0.478
NS	1.000	0.994	0.755	0.462
UL ($\alpha = 0$)	1.000	0.996	0.406	0.759
UL ($\alpha = 1$)	0.943	0.900	0.920	0.938
UL ($\alpha = 10$)	0.910	0.937	0.969	0.968

terms of *Certainty* and *Variety*. This study emphasizes the significance of examining the *Certainty* and *Variety* of each RGS.

5.4.3 Evaluation of *n*-best by various techniques

How to achieve high *Certainty* and *Variety*? One method to increase *Certainty* is to generate n-best lists with a wider range of responses, such that each n-best list is guaranteed to contain a specific number of noncontradictory responses. The diverse beam search (DBS) (Vijayakumar et al., 2016) and nucleus sampling (NS) (Holtzman et al., 2020) methods are used to construct such n-best lists. Furthermore, Li et al. (2020) recently proposed RGSs that use unlikelihood (UL) training to assign low probabilities to contradict responses. Using these RGSs to generate n-best lists will almost certainly enhance both *Certainty* and *Variety*. We assess the n-best lists generated using these three strategies to see how much these techniques enhance *Certainty* and *Variety* (n-best lists generated using DBS and NS, and n-best lists generated using beam search together with the UL training). Appendix 5.6.3 contains a description of the techniques used for this evaluation.

Result. Table 5.3 displays the *Certainty* and *Variety* of the 10-best lists generated using BS, DBS, NS, and UL.[6] The values of α show the degree of UL loss during fine-tuning. Here UL with $\alpha = 0$ used the RGS fine-tuned with maximum likelihood in the same training settings as those used for UL with $\alpha > 0$. Thus, note that comparing UL with $\alpha = 0$ and $\alpha > 0$ allows a fair comparison between likelihood and unlikelihood training. The results reveal the properties of the n-best lists obtained for the three techniques, as well as the extent to which the techniques increase *Certainty* and *Variety*. The *Certainty* obtained using the

[6]For the BS, DBS, and UL, we obtained the 10-best lists setting beam size to 10. For the NS, we got the 10-best lists by performing nucleus sampling ten times.

DBS and NS method reach 1.0 for significantly lower search sizes than that for the BS to attain a *Certainty* of 1.0; the *Variety* for CNTQ-type inputs are less than 0.5 (chance rate). Thus, using the DBS and NS methods efficiently improves *Certainty* compared with the results obtained using the beam search; nevertheless, the methods do not simultaneously attain high *Certainty* and *Variety*. However, the *Certainty* obtained using UL with $\alpha > 0$ are greater than those obtained using the BS, and this was accomplished while maintaining higher *Variety* than those obtained using the BS and UL with $\alpha = 0$ (likelihood training). Our findings show that RGSs are advancing toward high *Certainty* and *Variety*, which is particularly true for the recently proposed UL loss method. Despite the highly restricted viable responses, i.e., *yes* or *no*, the *Certainty* obtained using UL with $\alpha > 0$ does not reach 1.0. Thus, we conclude that there is still room for improvement in n-best list generation in terms of avoiding contradiction.

5.5 Conclusion

Based on the recent development of contradiction detectors, removing contradictory candidates from RGSs' n-best lists is a practical method for avoiding contradiction. In this method, the consistency of all candidates in the n-best lists substantially affects whether the final outputs are contradictory, i.e., practical CA.

We quantitatively examined the properties of the n-best lists in terms of avoiding contradiction, using polar-typed questions as inputs. We demonstrated that the proposed framework exhibits the properties of n-best lists based on *Certainty* and *Variety*. *Certainty* determines whether an n-best list has at least one noncontradictory response, whereas *Variety* evaluates how many noncontradictory responses each n-best list has. The results, particularly, demonstrated the present limitations on achieving high *Certainty* and *Variety* when using the well-established beam search method. In addition, our method emphasizes the improvements in *Certainty* and *Variety* achieved by recently proposed response generation strategies.

5.6 Appendix

5.6.1 Details of transforming NLI data

As described in Section 5.3.1, we obtain an input from the NLI dataset. Specifically, we convert the hypothesis sentence of an NLI sample into a yes-no question. We describe the procedure as follows:

1. Detect the first verb of a sentence.
2. Move the verb to the beginning of the sentence, or put one of {*Do*, *Does*, *Did*} at the front of the sentence, changing the verb back to its base (e.g., *made* → *make*).
3. Change first-person pronouns to second-person pronouns and second-person pronouns to first-person pronouns (e.g., *my* → *your*).
4. Change the punctuation mark at the end of the sentence to a question mark.

We used spaCy (`en_core_web_sm`) (Honnibal and Montani, 2017) to detect the verbs of hypothesis sentences. We did not use NLI samples with syntactically complex hypothesis sentences, such as those containing coordinating conjunctions, to avoid obtaining ungrammatical inputs. Further details are provided in our source codes.[7]

5.6.2 Details of yes-no classifier

Training settings. On the Circa dataset, we fine-tuned the pretrained RoBERTa (`roberta-large`) implemented by Hugging Face (Wolf et al., 2020). We divided the dataset at random into train : valid = 8 : 2. The other training parameters were identical to those used by Louis et al. (2020).

Performance of classifier. To investigate the performance of the classifier, we measured the classification accuracy. First, we manually labeled the top-1 responses in the 10-best lists generated by the evaluation presented in Section 5.4.2 with one of the two following labels: *Contradictory* or *Noncontradictory*. The accuracy with which the automated evaluation categorized the labeled responses was then evaluated. We selected 500 responses[8] from 50 EntQ/CntQ inputs drawn at random from our test for the evaluation. The classifier classified 433/500 responses (see Appendix 5.6.3), and the accuracy was 0.921. Some examples of the classification are shown in Table 5.4. The classifier correctly detected the contradiction in the RGS response using an indirect expression, in Example 1. However, in Example 2, the classifier failed to detect the contradiction of the RGS response, having both a noncontradictory direct expression ("No") and a contradictory indirect expression (the part of the response after "No"). We found that the classifier tended to misclassify RGS responses containing the contradictions with themselves, such as Example 2.

[7] https://github.com/shiki-sato/nbest-contradiction-analysis
[8] 100 responses generated by each of 5 RGSs.

Table 5.4 Examples of the response classification results by the yes-no classifier. The RGS responses were generated by Blender 400M using beam search with beam size $B = 10$.

(a) Example 1

History:	and we didn't ever call it uh Cokes and such you know we call it soda.
Message:	Don't you always call it Coke?
RGS Response:	We call it coke.
Human Label:	**Contradictory**
Predicted Label:	**Contradictory**

(b) Example 2

History:	The buying a house was the last thing that i wanted to do.
Message:	Weren't you desperate to buy a house?
RGS Response:	No, I just wanted to buy a house.
Human Label:	**Contradictory**
Predicted Label:	**Non-contradictory**

5.6.3 Details of experiments

Number of evaluated stimulus inputs. To simplify the evaluation, we omitted from Section 5.4 and Appendix 5.6.2 the inputs with one or more ambiguous responses in the n-best lists. We defined ambiguous responses as those that were not identified by the classifier as either affirmations or refutations.[9] Table 5.5 and Table 5.6 display the number of inputs from the total of $2,000$ EntQ/CntQ used for the two evaluations in Section 5.4.

RGS settings. In Section 5.4 experiments, we used DialoGPT (Zhang et al., 2020) and Blender (Roller et al., 2021) as RGSs. We used the codes of ParlAI (Miller et al., 2017) with its default settings, except for `beam_length_penalty`$= 0$ to generate responses.

Unlikelihood training settings. We used unlikelihood training with Blender 3B for the study of Section 5.4.3. To use the unlikelihood training proposed by Li et al. (2020), we require training data that includes the following three elements: input (here, history, and message), gold response, and negative response. These training samples were created by altering the NLI data with entailing and contradicting hypotheses.[10] Table 5.7 displays the

[9] Circa dataset has seven different labels such as "Yes" and "Probably/sometimes yes." We regard the responses classified into "In the middle" or "I am not sure" as ambiguous ones.

[10] Note that we did not use the identical NLI samples to synthesize EntQ/CntQ.

Table 5.5 Number of stimulus inputs evaluated to calculate the *Certainty* and *Variety* described in Table 5.2.

RGS	EntQ	CntQ
Blender 400M	1331 / 2000	1270 / 2000
Blender 1B	1413 / 2000	1316 / 2000
Blender 3B	1566 / 2000	1403 / 2000
DialoGPT 345M	1126 / 2000	924 / 2000
DialoGPT 762M	1044 / 2000	956 / 2000

Table 5.6 Number of stimulus inputs evaluated to calculate the *Certainty* and *Variety* described in Table 5.3.

RGS	EntQ	CntQ
BS	1566 / 2000	1403 / 2000
DBS	991 / 2000	882 / 2000
NS	818 / 2000	684 / 2000
UL ($\alpha = 0$)	1914 / 2000	1871 / 2000
UL ($\alpha = 1$)	1806 / 2000	1887 / 2000
UL ($\alpha = 10$)	1654 / 2000	1811 / 2000

original NLI data and the transformed training samples. One NLI data set yields four types of questions (PositiveQ1, PositiveQ2, NegativeQ1, and NegativeQ2). We synthesized 8,000 samples from 2,000 NLI data and randomly divided them into $\mathrm{train} : \mathrm{valid} = 9 : 1$. We tuned the learning rate $\{7.0 \times 10^{-4}, 7.0 \times 10^{-5}, 7.0 \times 10^{-6}, 7.0 \times 10^{-7}, 7.0 \times 10^{-8}\}$ and the number of warmup updates $\{50, 100\}$ for each $\alpha = \{0, 1, 10\}$ for training. The rest of the training parameters are identical to those used by Roller et al. (2021). It is worth noting that we only trained the RGSs marked as UL in Section 5.4.3 on these transformed data.

Table 5.7 Example of transforming (a) original NLI data to (b) training sample for UL. We synthesized four questions, i.e., PositiveQ1, PositiveQ2, NegativeQ1, and NegativeQ2, from each NLI sample.

(a) Original NLI data

Premise:	yeah i'm in North Carolina
Hypothesis – **Entailment**:	I'm in North Carolina.
Hypothesis – **Contradict**:	I'm in South Carolina.

(b) Training samples for UL

PositiveQ1

History:	Yeah I'm in North Carolina.
Message:	Are you in North Carolina?
Gold:	Yes, I'm in North Carolina.
Negative:	No, I'm in South Carolina.

PositiveQ2

History:	Yeah I'm in North Carolina.
Message:	Are you in South Carolina?
Gold:	No, I'm in North Carolina.
Negative:	Yes, I'm in South Carolina.

NegativeQ1

History:	Yeah I'm in North Carolina.
Message:	Aren't you in North Carolina?
Gold:	Yes, I'm in North Carolina.
Negative:	No, I'm in South Carolina.

NegativeQ2

History:	Yeah I'm in North Carolina.
Message:	Aren't you in South Carolina?
Gold:	No, I'm in North Carolina.
Negative:	Yes, I'm in South Carolina.

Chapter 6

Conclusion

This thesis aimed to explore and construct a highly effective and practical framework for the automatic evaluation of CA. In particular, this thesis addressed the following research issues:

- **Which approach should we employ for automatic CA evaluation?** As described in Chapter 1, there are two evaluation approaches for CA: probability-based and generation-based. While the assessment through probability-based automatic evaluation presents an advantage over generation-based methods by eliminating the need for a high-performance contradiction detector, the validity of this approach remains unexplored.

- **Can we realize effective generation-based CA evaluation?** If the efficacy of probability-based evaluation falls short, an alternative recourse lies in adopting a generation-based approach. The merit of this method lies in its capacity to directly assess the actual responses generated by the RGS. However, the existing contradiction detector's accuracy does not meet the requisite standards for a pragmatic automatic generation-based evaluation. To achieve a proficient automatic CA evaluation, enhancements in its performance are imperative.

- **Can we realize practical generation-based CA evaluation?** As discussed in Chapter 5, it is important to consider the n-best generation of RGS in contradiction suppression. Therefore, in addition to improving contradiction detectors, it is also imperative to evaluate the consistency of all candidates within the n-best list generated by RGS in the context of CA assessment. Despite this, traditional generation-based automatic evaluation methods concentrate solely on the consistency of the 1-best candidates, neglecting a comprehensive analysis of the features exhibited by the n-best candidates generated by RGS.

This thesis made noteworthy contributions that can be succinctly outlined as follows:

- **Confirmed that we need to employ generation-based evaluation for efficient RGS improvement.** Our experimental findings conclusively demonstrated that, at the instance level, there was no significant correlation between the outcomes from probability-based automatic evaluation (i.e., response selection task for this study) and those derived from human evaluation. It is noteworthy, however, that a certain level of correlation was observed when considering evaluations at the rougher system level.

- **Improved the accuracy of data-driven contradiction detectors.** We hypothesized that the scarcity of contradiction data actually generated by RGSs constitutes a barrier to enhancing the performance of contradiction detectors. Consequently, we collected an extensive dataset comprising contradictory responses generated by RGSs to serve as training data for data-driven contradiction detectors. Our empirical investigations illustrated that training detectors on our dataset resulted in enhanced accuracy in identifying contradictions.

- **Proposed an n-best-aware CA evaluation framework.** We suggested assessing CA by examining the consistency of n-best candidates generated by RGSs, assuming the post-processing, wherein an optimal contradiction detector can select non-contradictory candidates from the pool of n-best candidates. Our experimental results with this framework showed the properties of n-best lists, which could be influential in suppressing contradictions. For example, beam search has limitations in avoiding contradiction, and recent techniques, such as unlikelihood training, can help these situations.

Around 2015, research on neural response generation models began to flourish (Shang et al., 2015), and with the advancement of hardware and software, the performance of open-domain response generation has dramatically improved. Initially, issues such as dull response generation were a concern, but by the time this study began in 2019, the scalability of the models allowed for the generation of informative responses maintaining relevance to the context (Zhang et al., 2020). However, as the ability to generate responses with relevance to context improved, more advanced challenges related to semantic appropriateness became apparent. Specifically, challenges related to consistency, non-toxicity, and factual correctness, as shown in Figure 2.1, emerged as serious problems. As a result, the focus in the field shifted towards addressing challenges associated with higher-level perspectives compared to relevance. However, as of 2024, despite various efforts, it is challenging to claim that these issues have been fully resolved. Even large-scale generative models like ChatGPT, which

have addressed various challenges of traditional systems, struggle with these issues. This suggests that these challenges related to semantic appropriateness, unlike those related to relevance, cannot be simply overcome by scaling up the model. Instead, tailored solutions need to be developed for each problem.

In this context, this thesis focuses on improving CA. As mentioned in Chapter 2, CA is closely tied to the improvement of other unresolved challenges. However, the progress in enhancing CA is still in its early stages. In particular, attempting to improve it faced the obstacle of lacking an automatic evaluation framework for trial and error. Therefore, this thesis initiates the establishment of an automatic evaluation metric for CA.

Note that the scope of this thesis is to establish an effective and practical framework for automatically evaluating CA. The subsequent step involves exploring strategies for the effective improvement of CA. One viable approach entails leveraging our evaluation frameworks themselves to actively enhance CA capabilities. As an illustration, reinforcement training of RGSs could be implemented using automatic evaluation results as rewards.

It is also important to note that merely suppressing contradictions is insufficient for ensuring error-free RGSs; addressing other errors is also essential. As for representative domain-independent challenges, as listed in Figure 2.1, issues related to non-toxicity and factual correctness persist. In terms of non-toxicity, it is known that the technology foundation for CA improvement can be leveraged, thus suggesting the potential for enhancing non-toxicity using CA improvement techniques developed based on our evaluation framework. Regarding factual correctness, there are indications that it may improve alongside CA enhancements. Furthermore, the appropriateness of responses is not solely domain-independent. For instance, the appropriate behavior from the RGS varies significantly between dialogues aimed at encouraging users and those where the system should act as a listener to the user's narrative. Improvements in domain-specific appropriateness still have ample room for exploration.

References

Addlesee, A., Eshghi, A., and Konstas, I. (2019). Current Challenges in Spoken Dialogue Systems and Why They Are Critical for Those Living with Dementia. In *Proceedings of the Dialogue for Good workshop on Speech and Language Technology Serving Society*.

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). Towards a human-like open-domain chatbot. In *arXiv preprint arXiv:2001.09977*.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the 5th international conference on learning representations (ICLR)*.

Bao, S., He, H., Wang, F., Wu, H., Wang, H., Wu, W., Guo, Z., Liu, Z., and Xu, X. (2021). PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525.

Bao, S., He, H., Wang, F., Wu, H., Wang, H., Wu, W., Wu, Z., Guo, Z., Lu, H., Huang, X., Tian, X., Xu, X., Lin, Y., and Niu, Z.-Y. (2022). PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 107–118.

Craswell, N. (2009). Mean reciprocal rank. In LIU, L. and ÖZSU, M. T., editors, *Encyclopedia of database systems*, pages 1703–1703.

Fainberg, J., Krause, B., Dobre, M., Damonte, M., Kahembwe, E., Duma, D., Webber, B., and Fancellu, F. (2018). Talking to myself: self-dialogues as data for conversational agents. *arXiv preprint arXiv:1809.06641*.

Goldzycher, J., Preisig, M., Amrhein, C., and Schneider, G. (2023). Evaluating the Effectiveness of Natural Language Inference for Hate Speech Detection in Languages with Limited Labeled Data. In Chung, Y.-l., R\textbackslash"ottger, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 187–201.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, pages 1891–1895.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The Curious Case of Neural Text Degeneration. In *Proceedings of the eighth international conference on learning representations (ICLR)*.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Dai, W., Madotto, A., and Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Jurafsky, D., Shriberg, L., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Kann, K., Ebrahimi, A., Koh, J., Dudy, S., and Roncone, A. (2022). Open-domain Dialogue Generation: What We Can Do, Cannot Do, And Should Do Next. In Liu, B., Papangelis, A., Ultes, S., Rastogi, A., Chen, Y.-N., Spithourakis, G., Nouri, E., and Shi, W., editors, *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165. Association for Computational Linguistics.

Kim, H., Kim, B., and Kim, G. (2020). Will I sound like me? Improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 904–916.

Komeili, M., Shuster, K., and Weston, J. (2022). Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Kottur, S., Wang, X., and Carvalho, V. R. (2017). Exploring personalized neural conversational models. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17)*, pages 3728–3734.

Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. (2017). Edina: Building an open domain socialbot with self-dialogues. In *1st Proceedings of Alexa Prize*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (ACL)*, volume 1, pages 994–1003.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. (2020). Don' t say that! Making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th annual meeting of the association for computational linguistics (ACL)*, pages 4715–4728.

Li, W., Kong, J., Liao, B., and Cai, Y. (2022). Mitigating Contradictions in Dialogue Based on Contrastive Learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2781–2788.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Kondrak, G. and Watanabe, T., editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Li, Z., Zhang, J., Fei, Z., Feng, Y., and Zhou, J. (2021). Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency. In *Findings of the joint conference of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (ACL-IJCNLP)*, pages 1057–1067.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Litman, D., Young, S., Gales, M., Knill, K., Ottewell, K., van Dalen, R., and Vandyke, D. (2016). Towards Using Conversations with Spoken Dialogue Systems in the Automated Assessment of Non-Native Speakers of English. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–275.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv preprint arXiv:1907.11692*.

Louis, A., Roth, D., and Radlinski, F. (2020). "I'd rather just go to bed": Understanding Indirect Answers. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7411–7425.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (2017). Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In Koller, A., Skantze, G., Jurcicek, F., Araki, M., and Rose, C. P., editors, *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Martinovsky, B. and Traum, D. (2003). The error is the clue: breakdown in human-machine interaction. In *Proceedings of Error handling in spoken dialogue systems*, pages 11–16.

Miller, A. H., Feng, W., Fisch, A., Lu, J., Batra, D., Bordes, A., Parikh, D., and Weston, J. (2017). ParlAI: A dialog research software platform. In *Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP): System demonstrations*, pages 79–84.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *Proceedings of the ninth International Conference on Learning Representations*.

Mündler, N., He, J., Jenko, S., and Vechev, M. (2023). Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. In *arXiv preprint arXiv:2305.15852*.

Nie, Y., Williamson, M., Bansal, M., Kiela, D., and Weston, J. (2021). I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In *Proceedings of the 59th annual meeting of the association for computational linguistics (ACL)*, pages 1699–1713.

Noh, E.-J. (1998). Echo Questions: Metarepresentation and Pragmatic Enrichment. *Linguistics and Philosophy*, 21(6):603–628.

OpenAI (2023). GPT-4 Technical Report. In *arXiv preprint arXiv:2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *arXiv preprint arXiv:2203.02155*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT)*, pages 2227–2237.

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). Assigning Personality/Profile to a chatting machine for coherent conversation generation. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18)*, pages 4279–4285.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J. (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume (EACL)*, pages 300–325.

Shang, L., Lu, Z., and Li, H. (2015). Neural Responding Machine for Short-Text Conversation. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., Behrooz, M., Ngan, W., Poff, S., Goyal, N., Szlam, A., Boureau, Y.-L., Kambadur, M., and Weston, J. (2022). BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. In *arXiv preprint arXiv:2208.03188*.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. (2016). Diverse Beam Search for Improved Description of Complex Scenes. In *Proceedings of the 32nd AAAI conference on artificial intelligence (AAAI-18)*.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. (2020). Neural Text Generation With Unlikelihood Training. In *Proceedings of the eighth international conference on learning representations (ICLR)*.

Welleck, S., Weston, J., Szlam, A., and Cho, K. (2019). Dialogue natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pages 3731–3741.

Williams, A., Nangia, N., and Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies (NAACL-HLT)*, volume 1, pages 1112–1122.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP): System demonstrations*, pages 38–45.

Xu, J., Szlam, A., and Weston, J. (2022). Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th annual meeting of the association for computational linguistics (ACL)*, volume 1, pages 2204–2213.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. In *arXiv preprint arXiv:2205.01068*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics (ACL): System demonstrations*, pages 270–278.

List of Publications

Journal Papers (Refereed)

1. 佐藤志貴, 赤間怜奈, 大内啓樹, 鈴木潤, 乾健太郎. 負例を厳選した対話応答選択による対話応答生成システムの評価. 会誌「自然言語処理」, Vol.29, No.1, 2022.

International Conference Papers (Refereed)

1. (Under review) Shiki Sato, Reina Akama, Jun Suzuki and Kentaro Inui. A Large Collection of Model-generated Contradictory Responses for Consistency-aware Dialogue Systems. ACL 2024.

2. Shiki Sato, Reina Akama, Hiroki Ouchi, Ryoko Tokuhisa, Jun Suzuki and Kentaro Inui. N-best Response-based Analysis of Contradiction-awareness in Neural Response Generation Models. SIGDIAL 2022.

3. Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki and Kentaro Inui. Evaluating Dialogue Generation Systems via Response Selection. ACL 2020.

Awards

1. Mar.2023, 言語処理学会 2022 年度最優秀論文賞.

2. Mar.2021, 言語処理学会第 27 回年次大会委員特別賞.

3. Dec.2019, 人工知能学会言語・音声理解と対話処理研究会 (SLUD) 第 87 回研究会 (第 10 回対話システムシンポジウム) 若手萌芽ポスター賞受賞.

Other Publications (Not refereed)

1. 佐藤志貴, 赤間怜奈, 鈴木潤, 乾健太郎. Follow-up 質問による矛盾応答収集の提案. 言語処理学会第 29 回年次大会.

2. 佐藤志貴, 赤間怜奈, 大内啓樹, 鈴木潤, 乾健太郎. 対話システムの矛盾応答の生成に対する脆弱性の分析, 言語処理学会第 27 回年次大会.

3. 佐藤志貴, 赤間怜奈, 大内啓樹, 鈴木潤, 乾健太郎. 対話応答選択による対話応答生成モデルの評価, 言語処理学会第 26 回年次大会.

4. 佐藤志貴, 赤間怜奈, 大内啓樹, 鈴木潤, 乾健太郎. 負例を厳選した対話応答選択データセット構築の試みと分析, 人工知能学会言語・音声理解と対話処理研究会 (SLUD) 第 87 回研究会 (第 10 回対話システムシンポジウム).

5. 佐藤志貴, 赤間怜奈, 大内啓樹, 鈴木潤, 乾健太郎. 負例を厳選した対話応答選択データセットの構築, 第 14 回 NLP 若手の会シンポジウム (YANS).