Contrastive Response Pairs for Automatic Evaluation of Non-task-oriented Neural Conversational Models

Koshiro Okano
Doshisha University

Yu Suzuki
Doshisha University

Masaya Kawamura
Doshisha University

Tsuneo Kato
Doshisha University

Akihiro Tamura
Doshisha University

Jianming Wu
KDDI Research, Inc.

Abstract

Responses generated by neural conversational models (NCMs) for non-task-oriented systems are difficult to evaluate. We propose contrastive response pairs (CRPs) for automatically evaluating responses from non-task-oriented NCMs. We conducted an error analysis on responses generated by an encoder-decoder recurrent neural network (RNN) type NCM and created three types of CRPs corresponding to the three most frequent errors found in the analysis. Three NCMs of different response quality were objectively evaluated with the CRPs and compared to a subjective assessment. The correctness obtained by the three types of CRPs were consistent with the results of the subjective assessment.

1 Introduction

Non-task-oriented dialogue systems must generate responses based on dialogue contexts although possible responses are not limited to a few correct answers. Neural conversational models (NCMs), such as an encoder-decoder RNN with an attention mechanism (Bahdanau et al., 2014; Shang et al., 2015; Sordoni et al., 2015) and Transformer (Vaswani et al., 2017), generate fluent responses; however, an automatic evaluation of response quality in non-task-oriented NCMs has not been established yet. Reference-based evaluation indices such as BLEU have a low correlation with subjective scores because of the diversity of possible responses. To address this problem, there have been various proposals such as an index referencing a model response and taking into account the previous utterance of the interlocutor (Tao et al., 2017), an index integrating subjective and statistical evaluations (Hashimoto et al., 2019), and an interactive evaluation method assuming that the quality can only be evaluated through interaction (Ghandeharioun et al., 2019).

On the other hand, neural machine translation (NMT) has improved its quality at the sentence level, and context awareness (i.e., consistency between translated sentences when processing a text or series of sentences) still remains a challenge. Sennrich et al. proposed contrastive discourse sets to evaluate how well NMT models handle anaphoric pronouns, and coherence and cohesion for context-aware NMT (Bawden et al., 2018), by extending his proposed contrastive translation pairs (CTPs) (Sennrich, 2017). A CTP consists of a correct translation and an incorrect one in which a minimal number of words is substituted with wrong ones. The model quality is measured on correctness, i.e., the ratio of the number of pairs in which the correct translation received a higher score in forced decoding than the incorrect one to the total number of pairs. Voita et al. further analyzed errors in context-aware English-Russian NMT to extract frequent error patterns and proposed a set of CTPs to evaluate the accuracy of an NMT in terms of the frequent error patterns (Voita et al., 2019).

In this paper, we propose contrastive response pairs (CRPs) for automatically evaluating the quality of NCM responses with reference to the CTPs for evaluating context-aware NMT. We first conducted an error analysis on responses generated by NCMs trained on a large-scale conversation corpus. Then, we created a set of CRPs corresponding to three frequent error patterns. Finally, we examined whether the CRPs correctly reflected the difference in NCM response quality by comparing the correctness of the CRPs and the results of a subjective assessment on three NCMs with varying levels of quality. Specifically, we proceeded in the following steps.

1. Error Analysis: We conducted a binary classification of responses generated by NCMs in

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 202–207
July 29–31, 2021. ©2021 Association for Computational Linguistics

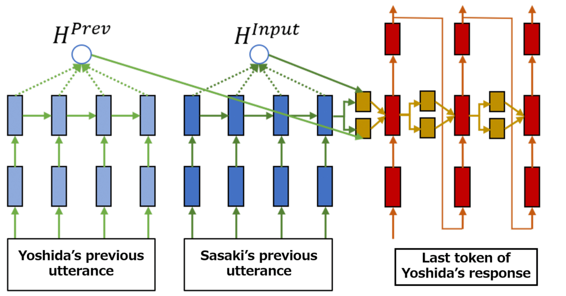

Figure 1: Architecture of double attention model.

terms of naturalness in the dialogue context. Then, we further classified the responses that were judged unnatural into 10 error classes manually and counted their frequencies.

2. Creation of CRPs: A set of CRPs was created by manually extracting contextually-correct responses from the conversation corpus, adding an error with minimal modification to every correct response, and pairing it with the correct response to form a CRP.

3. Model Evaluation: Forced decoding was conducted on the correct and incorrect responses of each CRP, and the correctness was measured. The correctness of the different models was compared to see if they are consistent with the results of the subjective assessment.

These three steps are discussed in the following sections in detail.

2 Error Analysis of Responses Generated by Neural Conversational Models

We simulated conversation between women using NCMs. We used a large-scale fictive conversation corpus between two Japanese ladies "Miss Yoshida" and "Miss Sasaki" for training and evaluating the NCMs. The corpus consists of 1.68 million fictive conversations compiled by 200 crowd-workers. The characters were kept consistent by specifying detailed personas across 80 items, which were shared among crowd-workers. We extracted 1.1M, 64k and 64k of Yoshida's utterances with preceding dialogue contexts for training, validation, and evaluation of Yoshida model.

We trained a GRU-based encoder-decoder RNN model with an attention mechanism, the network architecture of which is shown in Figure 1. The model received Yoshida's and Sasaki's previous utterances with two encoders, and output Yoshida's response. We refer to this model as the "Double attention model." The model was trained by teacher forcing with the cross-entropy loss function.

Table 1: Definition of ten error classes.

Label	Description
ICW	Containing contextually inappropriate content words
RUDE	Speaking rudely to interlocutor
FNC	Selecting inappropriate function words
ESE	Selecting inappropriate end-of-sentence expression
SC	Self-contradicting to one's own previous utterance
RP	Repeating one's own previous utterance
NA	Not answering interlocutor
DIS	Incomprehensible response
COL	Collision of content word's attribute to past utterances
ETC	Others

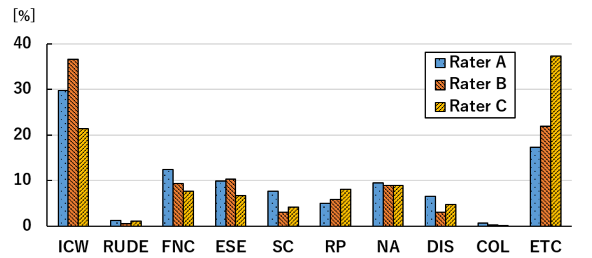

Figure 2: Relative frequency distribution of ten error classes labeled by three raters.

The double attention model generated responses on the basis of the maximum mutual information criterion (Li et al., 2016). We randomly sampled 3,000 responses from the validation set. Three of the authors manually analyzed errors in the 3,000 responses. First, they rated each response as natural or unnatural in its dialogue context. If it was unnatural, they determined the reason for unnaturalness using their own criteria. Then they negotiated with each other to unify the error classes and criteria. After the unity, they determined the reason for unnaturalness with the unified criteria for responses deemed unnatural by more than one rater. Table 1 lists the error classes, and Figure 2 shows the relative frequency distributions of the error classes labeled by the three raters.

On average, 41.9% of the responses were classified as unnatural. Cohen's kappa coefficients between all the pairs were 0.61. The unnatural responses were broken down into the distribution shown in Figure 2. The most frequent errors were caused by contextually-inappropriate content words (ICW, 28.9%), followed by inappropriate function words (FNC, 9.8%), inappropriate end-of-sentence expressions (ESE, 8.9%) and not answering the previous question (NA, 8.0%), not including others (ETC, 15.0%). We created CRPs to evaluate the performance of the NCM on the three

Table 2: Relative frequency distributions of subclasses in inappropriate end-of-sentence expression.

subclass	%
Switch between declarative and interrogative	33.3
Switch between affirmative and negative	11.1
Change of implicitly-meant subject	11.1
Missing empathic expression	8.9
Mischoice of tense	4.4
Mischoice of verb	4.4
Missing wishful expression	4.4
Others	22.2

most common errors, ICW, FNC and ESE.

3 Creation of Contrastive Response Pairs

3.1 CRP with Substituted Content Words

This CRP evaluates NCMs on selecting appropriate content words in terms of the dialogue context. To create a pair, we needed to select which content word to substitute, and what word to substitute it with. We processed the substitution semi-automatically. We manually selected a contextually-sensitive noun or compound noun to substitute, and examined two criteria to select a substitute word from a large vocabulary list.

Since it was not appropriate to select a linguistically unlikely substitute word, we trained a bigram language model and selected a substitute word on the basis of the following criteria: 1) A linguistic probability nearly equal to that of the original noun in the reference sentence (Equally-likely, EL), and 2) The highest linguistic probability (Most-likely, ML). When a word w_i in a sentence $W = \{w_1, \ldots, w_n\}$ is substituted with a word $\hat{w}_i$, the criteria were represented in equation (1) for EL and (2) for ML.

$$\hat{w}_i = \underset{v \in V}{\operatorname{argmin}} \left[\left\{ \log \frac{P(v|w_{i-1})}{P(w_i|w_{i-1})} \right\}^2 + \left\{ \log \frac{P(w_{i+1}|v)}{P(w_{i+1}|w_i)} \right\}^2 \right] \quad (1)$$

$$\hat{w}_i = \underset{v \in V}{\operatorname{argmax}} \left\{ \log P(v|w_{i-1}) + \log P(w_{i+1}|v) \right\} \quad (2)$$

Note that the vocabulary V consists of nouns appearing in the corpus more than once and excludes words included in the inputs into the encoders. Table 7 in Appendix shows an example of the contrastive response pair (ML) with a substituted content word.

3.2 CRP with Substituted End-of-Sentence Expression

Japanese is an agglutinative language, so the meaning of a sentence changes depending on its end-of-sentence expression. Affirmative or negative, declarative or interrogative, and other nuances are determined by the end-of-sentence expression. We further classified the ESE errors into subclasses manually. Table 2 shows the subclasses and their relative frequency distribution. The most frequent subclass was switching between declarative and interrogative, followed by switching between affirmative and negative, and changing an implicit subject due to an ESE error. Japanese is a null-object language; thus, a subject can be omitted from a sentence when it is obvious from context. An inappropriate ESE may change the implicit subject. Here, we omit details of the less frequent subclasses due to limitations in space.

We created CRPs corresponding to the two most frequent error subclasses "declarative and interrogative" and "affirmative and negative." We created the two types of CRPs manually on the basis of a simple rule that switch the two types of end-of-sentence expression randomly. Table 8 in Appendix shows an example of the CRP with a substituted end-of-sentence expression.

3.3 CRP with Substituted Function Words

Japanese has flexible word order, and function words, namely particles, determine the deep cases of content words. Incorrect use of function words results in unnaturalness and sometimes makes a sentence incomprehensible.

We created CRPs in which a particle was substituted with another particle. Since some particles are similar in meaning, we substituted particles randomly under the condition that they change the deep case of the content word. An example of CRPs with substitution of function words is listed in Table 9 in Appendix.

4 Evaluation

4.1 Experimental Setup: NCMs for Comparison and Subjective Assessment

We created a total of 1,160 CRPs: 350 pairs each for EL and ML for substituted content words, 270 pairs with substituted end-of-sentence expression, and 190 pairs with substituted function words.

We trained the following three NCMs each having a different performance level:

- Double attention: A model with two encoders, one decoder, and an attention for each encoder. The model was used in the error analysis in Section 2.

Table 3: Relative frequency distributions of subjective assessment scores on appropriateness of responses.

	1	2	3
No attention	27.4%	20.6%	52.0%
Single attention	26.6%	20.5%	53.0%
Double attention	23.3%	22.2%	54.5%

Table 4: Ratios of three error classes subjectively labeled on responses that were rated 1.

	a) ICW	b) ESE	c) FNC
No attention	22.5%	5.2%	2.9%
Single attention	22.0%	5.0%	3.3%
Double attention	19.5%	4.9%	4.4%

- Single attention: A model with an encoder, a decoder, and an attention for Sasaki's previous utterance. Yoshida's previous utterance cannot be taken into account.

- No attention: A model with an encoder for Sasaki's previous utterance and an decoder, but no attention.

Since the Single attention and No attention models were degraded models with respect to Double attention model, the quality of the generated responses was expected to be lower in the order of Double attention, Single attention and No attention. We conducted a crowdsourced subjective assessment to verify the order of the quality. The three NCMs generated responses for 1,200 dialogue contexts. The crowd-workers were instructed to assess the appropriateness of the responses on a 3-point scale: 1: inappropriate, 2: difficult to judge and 3: appropriate. Additionally, we asked them to check any of the following three boxes: a) inappropriate content word (ICW), b) inappropriate end-of-sentence expression (ESE), and c) inappropriate function word (FNC) if a response that they rated 1 falls into any of the error classes. Each response was assessed by five raters, resulting in 6,000 votes in total for each NCM.

Table 3 shows the relative frequency distribution of the subjective scores. The number of responses rated 3 increased and those rated 1 decreased in the order of No attention, Single attention and Double attention as expected.

Table 4 shows the ratios of the error classes subjectively labeled by the raters on the responses they rated 1 in Table 3. The ratios of ICW and ESE decreased in the order of No attention, Single attention, and Double attention, while the ratio of FNC increased in that order.

Table 5: Correctness of three models with whole set and subsets of contrastive response pairs.

	ALL	ICW (EL)	ICW (ML)	ESE	FNC
No attention	88.9%	94.8%	80.0%	90.0%	93.1%
Single attention	89.2%	96.2%	81.1%	89.2%	91.5%
Double attention	89.5%	94.5%	82.0%	92.6%	89.4%

4.2 Results of CRP Evaluation

The correctness of the models with the whole set and subsets of CRPs is shown in Table 5. The correctness with the whole set (ALL) increased in the order of No attention, Single attention, and Double attention. This result was consistent with the overall results of the subjective assessment, i.e., responses rated 3 increased and those rated 1 decreased in that order.

The correctness with the two subsets of ICW showed different results. The correctness with the subset of ICW(EL) was very high in general and inconsistent with the ratio of subjectively labeled ICW errors shown in Table 4. Meanwhile, the correctness with the subset of ICW (ML) was not very high and consistent with the results of subjectively labeled ICW errors. The results indicate that the subset of ICW (EL) was too easy for the NCMs to select the right answer, and the subset of ICW (ML) was better-suited for automatic evaluation.

The correctness with the subset of ESE increased in the order of Single attention, No attention and Double attention. The result was consistent with the results of subjectively labeled ESE errors in that Double attention was the most effective among the three, while it was partly inconsistent in that No attention surpassed Single attention. Lastly, the correctness with the subset of FNC decreased in the same order, which was consistent with the ratio of subjectively labeled FNC errors.

5 Conclusion

We proposed contrastive response pairs (CRPs) for automatically evaluating neural conversational models for non-task-oriented dialogue systems. Three types of CRPs were created on the basis of an error analysis of responses generated by NCMs, and their capability of measuring NCM performance was examined using three NCMs of varying quality. The correctness given by automatic evaluation with the CRPs was mostly consistent with the results of a subjective assessment. In future work, we will increase the size of CRPs and create CRPs automatically.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *arXiv preprint, arXiv:1409.0473*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL-HLT 2018*, pages 1304–1313.

Asma Ghandeharioun, Judy Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedrize, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Proceedings of NIPS 2019*, pages 13658–13669.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of NAACL 2019*, pages 1689–1701.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-prompting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*, pages 110–119.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of EACL 2017*, pages 376–382.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of ACL-IJCNLP 2015*, pages 1577–1586.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT 2015*, pages 196–205.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *arXiv preprint, arXiv:1701.03079*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS 2017*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of ACL 2019*, pages 1198–1212.

A Appendix

Table 6: Sample responses of ten error classes

ICW (Original in Japanese)	
Yoshida	団子も好きですなんか食べたくなってきちゃったなー
Sasaki	私もですコンビニに行ってこようかな
Yoshida	バイキングはどこにあるんですか？
ICW (Translation in English)	
Yoshida	I like dumplings, too. I feel like eating them.
Sasaki	So do I. I will buy one at a convenience store.
Yoshida	Where are the **buffet**?
RUDE (Original in Japanese)	
Yoshida	カフェの制服可愛いですよね
Sasaki	可愛い制服のカフェって憧れだったんです
Yoshida	メイドみたいですね
RUDE (Translation in English)	
Yoshida	Uniforms worn at cafés are cute, aren't they?
Sasaki	I've wanted to wear a cute uniform at a café.
Yoshida	You must look like a **maid**.
FNC (Original in Japanese)	
Yoshida	あっそうか免許が必要なんですね
Sasaki	そうなんですよ取りたいです
Yoshida	合宿免許でも取りたいですね
FNC (Translation in English)	
Yoshida	Ah. You need a driver's license.
Sasaki	Yes, I want to get my license.
Yoshida	It's good to get it **also** in a driver's license camp.
ESE (Original in Japanese)	
Yoshida	あの連続ポイントは息を飲みました
Sasaki	そうそう分かる人がいてくれて嬉しいです
Yoshida	興奮していますよ
ESE (Translation in English)	
Yoshida	Your scoring one after another took my breath away.
Sasaki	Oh, yes. I'm happy that you knew that.
Yoshida	I **am** so excited.
SC (Original in Japanese)	
Yoshida	お寿司はわさび入りの物を食べますよ
Sasaki	そうなんですねー
Yoshida	ワサビが苦手なんですよ
SC (Translation in English)	
Yoshida	I usually eat sushi with wasabi.
Sasaki	Oh, I see.
Yoshida	**I do not like wasabi.**
RP (Original in Japanese)	
Yoshida	おでんには何もつけませんか？
Sasaki	そうですね特につけてません
Yoshida	おでんには何ですか？
RP (Translation in English)	
Yoshida	Don't you put anything on oden?
Sasaki	No, nothing.
Yoshida	**What do you put on oden?**
NA (Original in Japanese)	
Yoshida	ポコポコとか面白いですよ
Sasaki	LINE のゲームでしたっけ？
Yoshida	ゆるい感じですね
NA (Translation in English)	
Yoshida	There are interesting games like Pokopoco.
Sasaki	Is it a game on LINE?
Yoshida	**It has a relaxing mood.**
DIS (Original in Japanese)	
Yoshida	同じく結婚してませんいい相手がいればしたいですね
Sasaki	吉田さんは好きな人はいるんですか？
Yoshida	いませんよ〜〜いますよ
DIS (Translation in English)	
Yoshida	I'm not married either, I would like to if I had a good partner.
Sasaki	Ms. Yoshida, do you have someone you like?
Yoshida	**No, I don't. Yes, I do.**
COL (Original in Japanese)	
Yoshida	どの温泉がお気に入りですか？
Sasaki	わたしは草津温泉ですね道後温泉はどのあたりがよかったんですか？
Yoshida	草津温泉が有名ですよ
COL (Translation in English)	
Yoshida	What is your favorite hot spring?
Sasaki	I like Kusatsu Onsen hot spring. What did you like about Dogo Onsen hot spring?
Yoshida	**Kusatsu Onsen** is famous.
ETC (Original in Japanese)	
Yoshida	どちらにしても勉強は苦手でしたけどね
Sasaki	そうですか得意不得意ありますからね
Yoshida	バタバタバタフライ
ETC (Translation in English)	
Yoshida	I wasn't very good at studying anyway, though.
Sasaki	Well, we all have strong and weak points.
Yoshida	**Butter butter butterfly.**

Table 7: Example of contrastive response pair with substituted content word (in translation)

Yoshida	I feel Japanese food is best-suited for me.
Sasaki	It's Japanese food that we can eat every day and never get tired of it.
Yoshida (reference)	What is your favorite ingredient for **miso soup**?
Yoshida (error)	What is your favorite ingredient for **holidays**?

Table 8: Example of contrastive response pair with substituted end-of-sentence expression (in translation)

Yoshida	I prefer curry in a sweet taste.
Sasaki	Are you weak in a hot curry?
Yoshida (reference)	**Yes, I am.**
Yoshida (error)	**Am I?**

Table 9: Example of contrastive response pair with substituted function word (in translation)

Yoshida	If you live on your own, you can probably enjoy cooking more.
Sasaki	It is probably true.
Yoshida (reference)	A lady **good at cooking** is popular with men, huh?
Yoshida (error)	A lady **who is cooked** is popular with men, huh?