Is a Knowledge-based Response Engaging?: An Analysis on Knowledge-Grounded Dialogue with Information Source Annotation

Takashi Kodama[1], **Hirokazu Kiyomaru**[1]
Yin Jou Huang[1], **Taro Okahisa**[2], **Sadao Kurohashi**[1,3]
[1]Kyoto University, [2]Shizuoka University, [3]National Institute of Informatics
{kodama, kiyomaru, huang, kuro}@nlp.ist.i.kyoto-u.ac.jp
okahisa-taro@inf.shizuoka.ac.jp

Abstract

Currently, most knowledge-grounded dialogue response generation models focus on reflecting given external knowledge. However, even when conveying external knowledge, humans integrate their own knowledge, experiences, and opinions with external knowledge to make their utterances engaging. In this study, we analyze such human behavior by annotating the utterances in an existing knowledge-grounded dialogue corpus. Each entity in the corpus is annotated with its information source, either derived from external knowledge (database-derived) or the speaker's own knowledge, experiences, and opinions (speaker-derived). Our analysis shows that the presence of speaker-derived information in the utterance improves dialogue engagingness. We also confirm that responses generated by an existing model, which is trained to reflect the given knowledge, cannot include speaker-derived information in responses as often as humans do.

1 Introduction

More and more dialogue research has utilized external knowledge to enable dialogue systems to generate rich and informative responses (Ghazvininejad et al., 2018; Zhou et al., 2018; Moghe et al., 2018; Dinan et al., 2019; Zhao et al., 2020). The major focus of such research is in how to select appropriate external knowledge and reflect it accurately in the response (Kim et al., 2020; Zhan et al., 2021; Rashkin et al., 2021; Li et al., 2022).

However, as shown in Figure 1[1], a good speaker not only informs the dialogue partner of external knowledge but also incorporates his or her own knowledge, experiences, and opinions effectively, which makes the dialogue more engaging. The extent to which models specializing in reflecting given external knowledge can achieve such an engaging behavior has not yet been explored quantitatively.

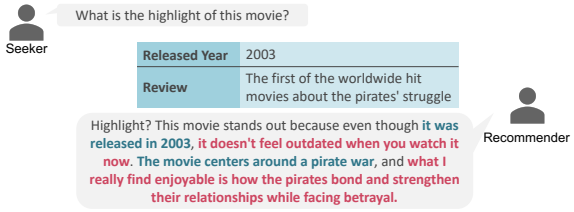

Figure 1: An example of Japanese Movie Recommendation Dialogue (Kodama et al., 2022). The table above the recommender's utterance indicates the external knowledge used in that utterance. The recommender incorporates not only database-derived information but also speaker-derived information.

In this study, we first analyze how humans incorporate speaker-derived information by annotating the utterances in an existing knowledge-grounded dialogue corpus. Each entity in the utterances is annotated with its information source, either derived from external knowledge (database-derived) or the speaker's own knowledge, experiences, and opinions (speaker-derived). The analysis of the annotated dataset showed that engaging utterances contained more speaker-derived information.

In addition, we train a BART-based response generation model in a standard way, i.e., by minimizing perplexity, and investigate the extent to which it incorporates speaker-derived information. The result showed that the response generation model did not incorporate speaker-derived information into their utterances as often as humans do. This result implies that minimizing perplexity is insufficient to increase engagingness in knowledge-grounded response generation and suggests room for improvement in the training framework.

[1]Examples of dialogues presented in this paper are originally in Japanese and were translated by the authors.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 237–243

2 Information Source Annotation

This section describes the annotation scheme for information sources and the annotation results.

2.1 Scheme

We annotate Japanese Movie Recommendation Dialogue (JMRD) (Kodama et al., 2022) with information sources[2]. JMRD is a human-to-human knowledge-grounded dialogue corpus in Japanese. A recommender recommends a movie to a seeker. Each utterance of the recommender is associated with movie information as external knowledge. Each piece of knowledge consists of a knowledge type (e.g., title) and the corresponding knowledge contents (e.g., "Marvel's The Avengers").

In this study, we extract entities from the recommender's utterances and annotate them with their information source. Entities are nouns, verbs, and adjectives and are extracted together with their modifiers to make it easier to grasp their meanings. Entities are extracted using Juman++ (Tolmachev et al., 2020), a widely-used Japanese morphological analyzer. Annotators classify the extracted entities into the following information source types:

Database-derived: The entity is based on the external knowledge used in that utterance.

Speaker-derived: The entity is based on the knowledge, experiences, and opinions that the recommender originally has about the recommended movie.

Other: The entity does not fall under the above two types (e.g., greetings).

An annotation example is shown below.

(1) Utterance: The action scenes$_{(\text{database})}$ are spectacular$_{(\text{speaker})}$!
Used knowledge: Genre, Action

We recruited professional annotators, who are native Japanese speakers, to annotate these information source types. One annotator was assigned to each dialogue. After the annotation, another annotator double-checked the contents.

2.2 Result

Table 1 shows the annotation statistics. While JMRD is a knowledge-grounded dialogue corpus and thus inherently contains many database-derived entities, it also contains about 60,000 speaker-derived entities. This result verifies that humans incorporate their own knowledge, experiences, and opinions into their utterances, even in dialogues to convey external knowledge.

[2]Examples of dialogue and knowledge in JMRD can be found in Appendix A.1.

	Train	Dev	Test	Total
# dialogues	4,575	200	300	5,075
# utterances (R)	51,080	2,244	3,347	56,671
# entities	235,771	10,320	15,734	261,825
# database-derived	166,958	7,223	10,476	184,657
# speaker-derived	51,170	2,303	4,095	57,568
# other	17,643	794	1,163	19,600

Table 1: Statistics of the information source annotation. R indicates recommender.

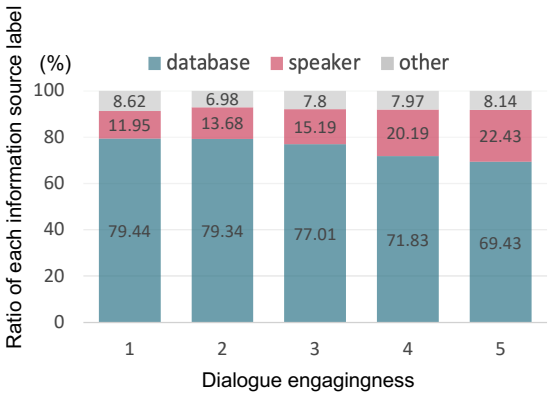

Figure 2: Relationship between dialogue engagingness and ratio of each information source label.

3 Analysis of Human Utterances

We analyze human utterances at the dialogue level and utterance level.

3.1 Dialogue-level Analysis

4,328 dialogues in JMRD have post-task questionnaires on 5-point Likert scale (5 is the best.) We regard the rating of the question to the seekers (i.e., Did you enjoy the dialogue?) as dialogue engagingness and analyze the relationship between this and the ratio of each information source label.

Figure 2 shows that dialogues with high engagingness scores tend to have more speaker-derived entities (or less database-derived) than those with low engagingness scores. When constructing JMRD, recommenders were given a certain amount of external knowledge and asked to use that knowledge to respond. However, recommenders highly rated by their dialogue partners incorporated not only the given external knowledge but also speaker-derived information to some extent in their dialogues.

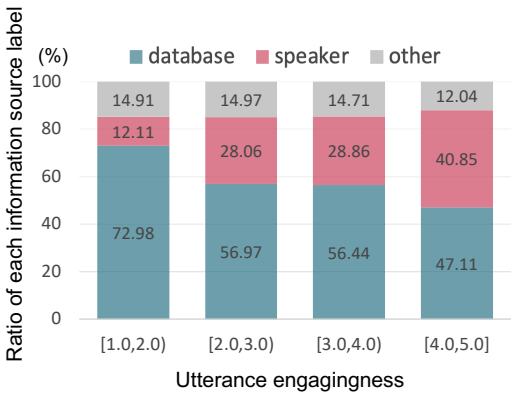

Figure 3: Relationship between utterance engagingness and ratio of each information source label.

3.2 Utterance-level Analysis

We conduct the utterance-level evaluation via crowdsourcing. We randomly extract 500 responses along with their contexts (= 4 previous utterances) from the test set. For each utterance, workers rate utterance engagingness (i.e., Would you like to talk to the person who made this response?) on a 5-point Likert scale, with 5 being the best. Three workers evaluate each utterance, and the scores are averaged.

The average score for utterances with speaker-derived entities was 3.31, while those without speaker-derived entities was 3.07. Student's t-test with p = 0.05 revealed a statistically significant difference between these scores.

Furthermore, Figure 3 shows the relationship between utterance engagingness and the ratio of each information source label. This figure shows that utterances with high scores tend to have more speaker-derived entities. This trend is consistent with that of the dialogue engagingness.

Does subjective knowledge contribute to engagingness? The knowledge type used in JMRD can be divided into subjective knowledge (review) and objective knowledge (title, etc.). Reviews are the opinions of individuals who have watched movies and have similar characteristics to speaker-derived information. We then examine whether there is a difference in engagingness between utterances using subjective and objective knowledge. The average engagingness scores were 3.32 and 3.16[3], respectively, and Student's t-test with p = 0.05 revealed no statistically significant difference. The above analysis demonstrates that information obtained from the speaker's own experience is an important factor in utterance engagingness.

[3]We exclude utterances referring to both of subjective and objective knowledge from this result.

4 Analysis of System Utterances

We investigate the distribution of information source labels in the responses of the model trained on the knowledge-grounded dialogue dataset. First, we train a Response Generator (§4.1) with the dialogue contexts and external knowledge as input and responses as output. Next, an Information Source Classifier (§4.2) is trained with responses and external knowledge as input and information source labels as output. Then, the Information Source Classifier infers the information source labels for the system responses generated by the Response Generator. Finally, we analyze the distribution of inferred information source labels.

4.1 Response Generator

We use a BART$_{large}$ (Lewis et al., 2020) model as a backbone.[4] The input to the model is formed as follows:

$$[CLS]u_{t-4}[SEP]u_{t-3}[SEP]u_{t-2}[SEP] \\ u_{t-1}[SEP][CLS_K]kt^1[SEP]kc^1[SEP]... \\ [CLS_K]kt^M[SEP]kc^M[SEP], \quad (1)$$

where t is the dialogue turn, u_t is the t-th response, and kt^i and kc^i $(1 <= i <= M)$ are the knowledge type and knowledge content associated with the target response, respectively (M is the maximum number of knowledge associated with u_t.) $[CLS_K]$ is a special token. We feed the gold knowledge into the model to focus on how knowledge is reflected in the responses. The model learns to minimize perplexity in generating u_t.

We evaluated the quality of response generation with the SacreBLEU (Post, 2018). BLEU-1/2/3/4 scored high, 81.1/73.5/71.0/69.9. This result is reasonable because the gold knowledge was given.

4.2 Information Source Classifier

We fine-tune a RoBERTa$_{large}$ (Liu et al., 2019) model.[5] The Information Source Classifier performs a sequence labeling task to estimate BIO[6]

[4]https://nlp.ist.i.kyoto-u.ac.jp/?BART%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB

[5]https://huggingface.co/nlp-waseda/roberta-large-japanese-seq512

[6]B, I and O stand for Begin, Inside and Outside, respectively.

		engagingness
Context	... Recommender: This movie is an animation movie released in 2015. Seeker: I see.	
Knowledge	{director, Takahiko Kyogoku}, {cast, Emi Nitta}, {cast, Yoshino Nanjo}	
Response	Human: The director is Takahiko Kyogoku, and the voice actors are Emi Nitta and Yoshino Nanjo. These two are also singers.	4.00
	System: The director is Takahiko Kyogoku. The voice actors are Emi Nitta and Yoshino Nanjo.	2.33

Table 2: An example of the human and system response. The blue and red parts refer to database-derived and speaker-derived information, respectively.

	Prec.	Rec.	F1
database-derived	94.92	95.61	95.27
speaker-derived	80.88	84.39	82.60
other	82.93	64.15	72.34
micro avg.	90.52	90.48	90.50

Table 3: Results of the sequence labeling by Information Source Classifier.

Dist. (%)	Human (gold)	Human (pred)	System (pred)
database-derived	66.22	66.75	85.48
speaker-derived	26.33	27.49	10.66
other	7.45	5.77	3.86

Table 4: Distributions of information source labels for human and system responses.

labels of the information source. The input to the model is formed as follows:

$$[CLS]u_t[SEP][CLS_K]kt^1[SEP]kc^1[SEP]... \\ [CLS_K]kt^M[SEP]kc^M[SEP] \quad (2)$$

Table 3 shows precision, recall, and F1 scores for each label and micro average scores across all labels. The micro average F1 score was 90.50, which is accurate enough for the further analysis.

4.3 Analysis for Inferred Labels

The information source labels for system responses are inferred using the classifier trained in Section 4.2. Table 4 shows distributions of information source labels for human and system responses. For a fair comparison, the human responses are also given labels inferred by the classifier (denoted as **Human (pred)**), although they have gold labels (denoted as **Human (gold)**). **Human (gold)** and **Human (pred)** have similar distributions, indicating that the accuracy of the classifier is sufficiently high. For **System (pred)**, the percentage of database-derived labels increased significantly (66.75%→85.48%) and that of speaker-derived information decreased significantly (27.49%→10.66%). This result shows that the response generation model, trained in a standard way, was not able to use speaker-derived information as often as humans do.

Ratio (%)	Human (gold)	Human (pred)	System (pred)
Title	30.21	34.12	27.09
Released Year	16.41	22.31	6.56
Director	13.94	11.96	4.50
Cast	36.11	45.34	23.45
Genre	10.47	15.14	5.49
Review	27.72	31.42	6.32
Plot	13.98	13.68	2.32
No knowledge	57.49	63.08	55.99

Table 5: Average ratios of speaker-derived labels per knowledge type used.

Table 2 shows an example of human and system responses along with the engagingness scores. The system was able to reflect given knowledge in the response appropriately but did not incorporate additional speaker-derived information, such as the information two voice actors also work as singers.

For further analysis, we investigated the average ratios of speaker-derived information by knowledge type used. Table 5 shows the result. Significant drops were observed for reviews (31.42%→6.32%) and plots (13.68%→2.32%). This is probably because reviews and plots are relatively long and informative external knowledge, so the system judged there was no need to incorporate additional speaker-derived information.

Combined with our observation that speaker-derived information improves engagingness, the current model is likely to have lower engagingness due to its inability to effectively incorporate speaker-derived information. Such an ability is hardly learned by simply optimizing a model to reduce the perplexity of response generation, suggesting the need for a novel learning framework.

5 Conclusion

We analyzed the distribution of speaker-derived information in human and system responses in the knowledge-grounded dialogue. The analysis showed that the use of speaker-derived information, as well as external knowledge, made responses more engaging. We also confirmed that the response generation model trained in a standard way generated less speaker-derived information than humans.

It is difficult to make good use of speaker-derived information by simply minimizing the perplexity of the model because a wide variety of speaker-derived information appears in each dialogue. We hope our published annotated corpus becomes a good launch pad for tackling this issue.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments. This work was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation. This work was also supported by JST, CREST Grant Number JPMJCR20D2, Japan and JSPS KAKENHI Grant Number JP22J15317.

References

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In International Conference on Learning Representations.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In International Conference on Learning Representations.

Takashi Kodama, Ribeka Tanaka, and Sadao Kurohashi. 2022. Construction of hierarchical structured knowledge-based recommendation dialogue dataset and dialogue system. In Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 83–92, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2810–2823, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 704–718, Online. Association for Computational Linguistics.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2020. Design and structure of the Juman++ morphological analyzer toolkit. Journal of Natural Language Processing, 27(1):89–132.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5621–5630, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3377–3390, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

A.1 Example of JMRD

Table 6 and 7 show examples of the dialogue and knowledge in JMRD.

A.2 Implementation Details

A.2.1 Response Generator

Dialogue contexts, knowledge (knowledge types and contents), and target responses are truncated to the maximum input length of 256, 256, and 128, respectively. The model is trained for up to 50 epochs with a batch size of 512 and 0.5 gradient clipping. We apply early stopping if no improvement of the loss for the development set is observed for three consecutive epochs. We use AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and an initial learning rate $= 1e-5$. We use an inverse square root learning rate scheduler with the first 1,000 steps allocated for warmup. During decoding, we use the beam search with a beam size of 3.

A.2.2 Information Source Classifier

Target responses and knowledge (knowledge types and contents) are truncated to the maximum input length of 128 and 384, respectively. The model is trained for up to 20 epochs with a batch size of 64 and 0.5 gradient clipping. We apply early stopping if no improvement of the f1 score for the development set is observed for three consecutive epochs. We use AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and an initial learning rate $= 1e-5$. We use an inverse square root learning rate scheduler with the first 1,000 steps allocated for warmup.

Turn	Dialogue	Knowledge type	Knowledge content
R_1	Hello.	No knowledge	-
S_1	Hello. Nice to meet you!		
R_2	Do you know “Avengers: Endgame”?	Title	Avengers: Endgame
S_2	I have only heard of the title...		
R_3	This movie was released in 2019.	Released Year	2019
S_3	Got it. Is it an American movie?		
R_4	Yes, It’s an American action movie.	Genre	Action
S_4	What are some of the highlights?		
R_5	The highlight is when the heroes gather to confront Thanos, who is an alien villain!	Review	Heroes gather to confront Thanos
S_5	I see! Is this a story of battles in space?		
R_6	No, it takes place on Earth.	No knowledge	-
S_6	Then, the villain will attack the earth...		
R_7	Yes, there are some scary moments.	No knowledge	-
S_7	Is it scary...? I don’t really like horror movies, but I like action ones. Would I be able to enjoy watching it?		
R_8	It is not scary like horror movies, so I think you will enjoy watching it!	No knowledge	-
S_8	Good! The fight between Thanos and the heroes sounds exciting!		
R_9	Please watch it!	No knowledge	-
S_9	Yes! I’ll have a chance to go to the video store soon and rent “Avengers: Endgame”!		
R_{10}	Thank you!	No knowledge	-
S_{10}	Thank you, too, for this valuable information!		

Table 6: A full dialogue example in JMRD. R and S in Turn column denote recommender and seeker, respectively. Subscript numbers indicate the number of turns in the dialogue. “No knowledge” means that the recommender did not use the given knowledge information.

Knowledge type		Knowledge content
Title		Avengers: Endgame
Released Year		2019
Director	name	Anthony Russo, Joe Russo
	description	Director, producer, screenwriter, actor, and editor for television and film in the United States.
Cast	$cast_1$ name	Robert Downey Jr.
	$cast_1$ description	an American actor, voice actor, musician, and producer.
	$cast_2$ name	Chris Evans
	$cast_2$ description	an American actor. He was born in Sudbury, Massachusetts.
Genre		Action, Adventure
Review		5 sentences, such as “Heroes gather to confront Samus.”
Plot		10 sentences, such as “In 2018, three weeks after half of all life in the entire universe was erased by decimation (genocide using the power of the Infinity Stone) by Thanos the Titan.”

Table 7: An example of knowledge used in JMRD. The director and the casts have two attributes: name and description, respectively.