Think Before You Speak: Learning to Generate Implicit Knowledge for Response Generation by Self-Talk

Pei Zhou[1]* Behnam Hedayatnia[2] Karthik Gopalakrishnan[2] Seokhwan Kim[2]
Jay Pujara[1] Xiang Ren[1] Yang Liu[2] Dilek Hakkani-Tur[2]
[1] Department of Computer Science, University of Southern California
[2] Amazon Alexa AI
{peiz,jpujara,xiangren}@usc.edu,
{behnam,karthgop,seokhwk,yangliud,hakkanit}@amazon.com

1 Introduction

End-to-end response generation (RG) models based on pre-trained transformer-based (Vaswani et al., 2017) language models (LM) have shown to produce human-like responses (Zhang et al., 2020; Lewis et al., 2020; Roller et al., 2020). Humans, however, not only "*just utter the right sentence*", but also contribute to the common ground consisting of mutual beliefs and common knowledge "*whose truth is taken for granted as part of the background of the conversation*" (Stalnaker, 1978; Clark and Schaefer, 1989; Clark and Brennan, 1991). For example, consider the utterance "*I need to buy some flowers for my wife*", a potential appropriate response is "*Perhaps you'd be interested in red roses*". To produce this response, the participant needs to understand relevant background knowledge such as "*rose is a type of flower*". Note that this background knowledge is *implicit* in the dialogue turns, meaning that an end-to-end RG system that takes utterances as input and outputs responses will omit this intermediate process that is crucial for humans in communications.

To fill the gap between current model's RG process and how humans use implicit background knowledge in conversations, we focus on *commonsense inference* as a type of background knowledge and aim to *train RG models that first generate implicit commonsense inferences before producing a response given previous utterances.* Inspired by inquiry-based discovery learning (Bruner, 1961) and the self-talk procedure (Shwartz et al., 2020), we encourage the RG model to talk with itself to elicit implicit knowledge before making a response.

Collecting training data is challenging for our purpose since commonsense knowledge is by definition often omitted in conversations. We use ConceptNet triples (Speer et al., 2017) as the knowledge schema to represent knowledge and align common sense-focused dialogues (Zhou et al., 2021) with knowledge to train RG models. We conduct extensive human evaluation on different variants of our training procedure and find that models that generate implicit background knowledge before responding produce more grammatical, coherent, and engaging responses compared to RG models that directly generate responses. Further analysis shows that models can sometimes even learn to distinguish *when it is necessary to self-talk to generate implicit knowledge*, i.e., be aware of potential knowledge gaps in dialogues. We hope our findings encourage more future studies on making RG models better emulate human communication process and produce better-quality responses.

2 Task Formulation and Setup

Given a dialogue history (multi-turn or single-turn) U, the task is to generate implicit background knowledge I and then produce a response R given both U and I. Extending most neural RG models that treat this as a *conditional language modeling* problem, we aim to learn the conditional probability distribution $P(I, R|U)$ by training on human dialogues.

2.1 Data Preparation

Knowledge Schema We consider *ConceptNet* (Speer et al., 2017) as our knowledge schema, which is a large-scale crowdsourced commonsense knowledge base consisting of triples such as "*buy, RelatedTo, money*". We have explored several other sources such as LMs trained to generate knowledge (Hwang et al., 2021; Becker et al., 2021) but observe much noise while aligning knowledge to dialogue turns.

Dialogues We use dialogue datasets from Zhou et al. (2021) as they propose "*common sense-focused dialogues*" by filtering three existing di-

[*] Work done while Pei Zhou was an intern at Amazon Alexa AI

Proceedings of the Third Workshop on Natural Language Processing for Conversational AI, pages 251–253
November 10, 2021. ©2021 Association for Computational Linguistics

alogue datasets DailyDialog (Li et al., 2017), EmpatheticDialogues (Rashkin et al., 2019), MuTual (Cui et al., 2020) using ConceptNet triples, and also crowdsourced SocialIQA-prompted dialogues (Zhou et al., 2021). We extract dialogues that each has at least one matched triple from ConceptNet in one of its consecutive turn pairs, resulting in around 31k dialogues and 159k turns we can use for training instances (excluding the first turn). The total number of turn pairs that have at least one matched triple is around 57k.

2.2 Training Setup

Each of our training instance is a sequence of tokens with three components: a dialog history U, implicit knowledge (empty or non-empty) I, and a response R. We enclose the implicit knowledge I with special symbols "<implicit>" and "< \implicit>" and add it between U and R. For example: "*<bos> <speaker1> I need to buy some flowers for my wife. <implicit> rose is a type of flower < \implicit> <speaker2> Perhaps you'd be interested in red roses <eos>*". We train the models to generate everything after <implicit> till <eos>, i.e. generate I and R conditioned on U. We use GPT2-medium (Radford et al.) and train the model for 3 epochs with batch size 4 and learning rate 6.25e-5.

Training Variations We also explore different ways to better train a model that can both self-talk and respond in dialogues. 1). **Two-Step training** Instead of jointly modeling $P(I, R|U)$ in a one-shot fashion, we consider split the instance to first train the model to generate I only based on U and then train it to generate R based on both I and U, i.e. $P(I|U)$ and then $P(R|U, I)$. 2). **Learning to distinguish when knowledge is needed.** In addition to using only the 57k instances where there is always non-empty implicit knowledge, we also evaluate using all dialogue instances, some of which have empty implicit knowledge. This allows us to evaluate if models can determine when to self-talk and when it is not needed (e.g., for simple chitchat utterances). 3). **Imbalanced VS balanced training set** Since we have more instances with empty implicit knowledge than non-empty knowledge (102k vs. 57k), we consider an imbalanced training set where we include all 159k instances and a balanced version where we randomly sample the same number of empty knowledge instances as non-empty ones (114k total).

	Grammatical	Coherent	Engaging
Prefers Self-Talk	**134**	**152**	**148**
Prefers Vanilla	116	121	120
Not Sure	50	27	32

Table 1: Human evaluation on three criteria by comparing vanilla RG model versus the *imbalanced* version

	Grammatical	Coherent	Engaging
Prefers Balanced	**145**	**139**	**154**
Prefers Imbalanced	128	128	128
Not Sure	27	33	18

Table 2: Human evaluation on three criteria by comparing self-talk models trained on *balanced* data versus the *imbalanced* version

3 Preliminary Results

For evaluation, we randomly sample 300 instances from unseen test dialogues for human evaluation. We provide a dialogue history and two model responses and ask three workers on Amazon Mechanical Turk (AMT) platform to select which one is better or not sure. We consider three criteria following most previous RG evaluation: which response is more *grammatical*, *coherent*, and *engaging*.

We generally find trained models follow the template of first generating implicity knowledge and then respond. Some example outputs to the utterance "*I need to buy some flowers for my wife*" are "*(buy is related to expensive) Can you afford the expensive flowers?*" and "*(buy is related to money) How much money do you want to spend?*".

Self-Talk models produce higher quality responses than vanilla RG models Table 1 shows that when comparing with a vanilla RG trained on dialogues, the two-step self-talk models improve on all three aspects, especially on coherence and engagingness.

Balanced training data results in better model Table 2 shows that although imbalanced self-talk model contains 45k more training instances than the balanced version, models trained on balanced data perform better on all three criteria, posing interesting phenomenon for later studies.

Due to page limit, we will summarize results for other model comparisons. We find two-step training outperforms one-step and learning to distinguish outperforms always outputs some implicit knowledge. We also find interesting new knowledge generated from models (not included in ConceptNet we used), e.g. "*bus is located at airport*".

References

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24.

Jerome S Bruner. 1961. The act of discovery. *Harvard educational review*.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Robert C Stalnaker. 1978. Assertion. In *Pragmatics*, pages 315–332. Brill.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.