Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations

Yang Deng[1], **Wenxuan Zhang**[2,†], **Yifei Yuan**[1], **Wai Lam**[1]
[1] The Chinese University of Hong Kong, [2] DAMO Academy, Alibaba Group
{dengyang17dydy,isakzhang}@gmail.com
{yfyuan,wlam}@se.cuhk.edu.hk

ACL 2023

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

Abstract

- Unlike empathetic dialogues,

  emotional support conversations (ESC)

  : help-seeker를 위로하기 위한 empathy를 전달하는 것 + 그들의 문제를 exploring하고 addressing할 수 있게 assist 해야 함

- 이 논문에서는

  - mixed initiative ESC의 문제점을 분석하고 발화를 speaker role과 initiative type으로 분류하는 tailor-designed schema 제안

  - mixed-initiative interactions을 측정할 수 있는 emotional support metrics 제안

  - mixed-initiative responses생성을 위한 knowledge-enhanced mixed initiative framework (KEMI) 제안

  - mixed initiative란

    : user와 system 모두 interaction direction을 주도하는 주도권 (initiative)를 가지는 시스템

Introduction

- As the world is making efforts to recover from Covid-19,
  emotional support가 emotional distress와 psychiatric illness 해결에 매우 중요한 역할

- 그래서 emotional support conversation (ESC)에 대한 연구들이 활발히 진행
  ESC 시스템은 user의 <u>emotional distress를 줄이는 것</u>과 동시에 대화를 통해서 <u>user의 문제를 파악하고 극복</u>하도록 돕는 것을 목표로 해야 함

Introduction

- different interaction patterns between **ESC** and **empathetic dialogues (ED)**

  : ED는 일반적으로 passive role인 반면, ESC는 대화 중에 능동적으로 initiative role을 바꾸면서 대화

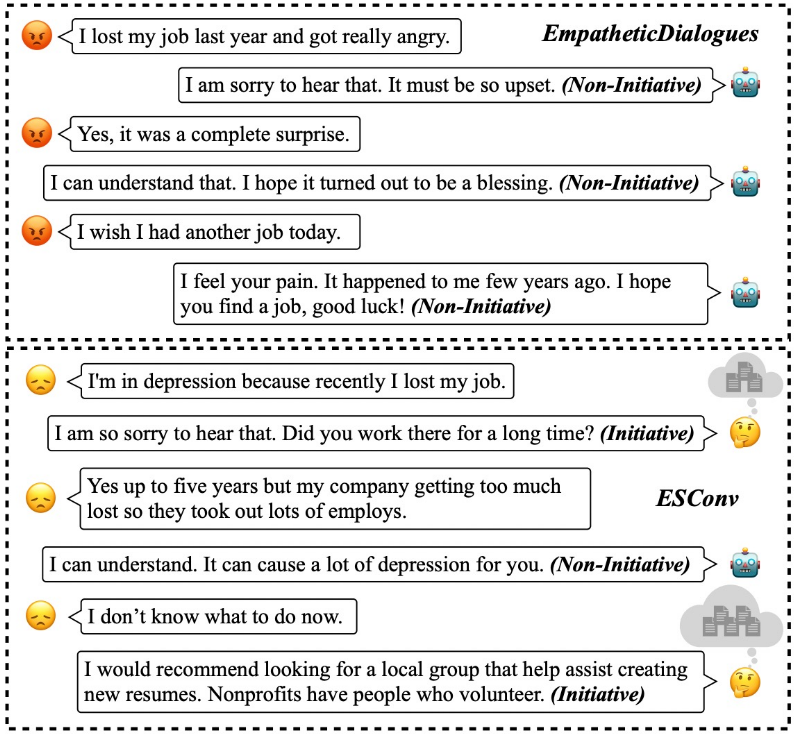


Figure 1: Examples from EMPATHETICDIALOGUES and ESConv datasets with a similar job loss problem.

ED는 user의 감정과 상황을 보고 사용자를 위로하는 것, 즉 Non-initiative를 목표로 함

ESC는 질문을 통해서 user의 문제를 적극적으로 탐색하고, useful information or supportive suggestion을 통해서 user가 문제를 극복할 수 있게끔

Introduction

- Mixed initiative

  : intrinsic feature of human-AI interactions

  : user와 system 모두 interaction direction을 주도하는 주도권 (initiative)를 가질 수 있는 것

- 이 두가지 개념이 더해진 mixed initiative ESC system 필요

  : 적절하게 empathetic response를 생성하거나 problem-solving discussion을 위해 대화 주도권을 switch

- 기존에는 각각에 대한 연구는 있었지만 두 개념을 합친 연구는 없었음

Introduction

- Mixed initiative ESC system

    1) When should the system take the initiative during the conversation?

        : 대화를 진행하면서 언제 주도권을 가져와야 하는지

    2) What kind of information is required for the system to initiate a sub dialogue?

        : 대화를 주도하려면 어떤 정보가 필요한지

    3) How could the system facilitate the mixed-initiative interactions?

→ 이러한 challenge를 고려한 framework를 제안

Introduction

- EAFR schema 제안

  annotate the utterances into different types with **speaker roles** and **initiative types**

- four emotional support metrics 제안

- a novel framework, named Knowledge Enhanced Mixed-Initiative model (KEMI) 제안

  external domain-specific knowledge 사용

  → Mixed initiative가 ESC에서 중요하다는 것 증명

  → Mixed initiative에서도 기존의 방법론보다 뛰어난 성능을 보이는 것을 증명

Preliminary Analysis

EAFR Schema & Metrics

- speaker roles과 initiative type에 따라서 4가지로 구분

  : **E**xpression (User-initiative), **A**ction (System-initiative), **F**eedback (User Non-Initiative), **R**eflection (System Non-Initiative)

Role	Type	EAFR	Definition	Sample Utterances
User	Initiative	Expression	The user describes details or expresses feelings about the situation.	My school was closed due to the pandemic. I feel so frustrated.
System	Initiative	Action	The system requests for information related to the problem or provides suggestions and information for helping the user solve the problem.	How are your feelings at that time? Deep breaths can help people calm down. Some researches has found that ...
User	Non-Initiative	Feedback	The user responds to the system's request or delivers opinions on the system's statement.	Okay, this makes me feel better. No, I haven't.
System	Non-Initiative	Reflection	The system conveys the empathy to the user's emotion or shares similar experiences and feelings to comfort the user.	I understand you. I would also have been really frustrated if that happened to me. I'm sorry to hear about that.

Table 1: Definition and Examples for EAFR Schema Reflecting Patterns of Initiative Switch between Dialogue Participants in Emotional Support Conversations.

Preliminary Analysis

EAFR Schema & Metrics

- four emotional support metrics

  each utterance i in a dialogue is annotated as a tuple (r_i, t_i, v_i, e_i)

  $r_i \in$ {User(U), System(S)} → Speaker role

  $t_i \in$ {Initiative(I), Non-Initiative(N)} → initiative role

  $v_i \in \{0, 1\}^{|V|}$ → One hot vocabulary embeddings

  $e_i \in [1, 5]$ → level of emotion intensity

Preliminary Analysis

EAFR Schema & Metrics

each utterance i in a dialogue is annotated as a tuple (r_i, t_i, v_i, e_i)

$r_i \in$ {User(U), System(S)} → Speaker role

$t_i \in$ {Initiative(I), Non-Initiative(N)} → initiative role

$v_i \in \{0, 1\}^{|V|}$ → One hot vocabulary embeddings

$e_i \in [1, 5]$ → level of emotion intensity

- four emotional support metrics

1) **Proactivity**: how proactive is the system in the emotional support conversation? (시스템이 얼마나 적극적인지)

$$\text{Pro} = \frac{1}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)} \sum_{i=1}^{n} \mathcal{I}(r_i = S, t_i = I)$$

→ system-initiative interactions의 비율

2) **Information**: how much information does the system contribute to the dialogue?

$$\text{Inf} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{|V|} \mathcal{I}(r_i = S, v_{ik} = 1, \sum_{j=1}^{i-1} v_{jk} = 0)}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)}$$

→ new frequent terms의 평균

→ 얼마나 이전 발화에 등장하지 않은 새로운 단어가 등장하는지

Preliminary Analysis

EAFR Schema & Metrics

each utterance i in a dialogue is annotated as a tuple (r_i, t_i, v_i, e_i)

$r_i \in$ {User(U), System(S)} → Speaker role

$t_i \in$ {Initiative(I), Non-Initiative(N)} → initiative role

$v_i \in \{0, 1\}^{|V|}$ → One hot vocabulary embeddings

$e_i \in [1, 5]$ → level of emotion intensity

- four emotional support metrics

3) **Repetition**: how often does the system follow up on the topic introduced by the user?

$$\text{Rep} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{|V|} \mathcal{I}(r_i = S, v_{ik} = 1, \sum_{j=1}^{i-1} v_{jk}[r_j = U] > 0)}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)}$$

→ System 발화 중에서, user가 언급한 단어를 얼마나 반복적으로 언급하는지

4) **Relaxation**: how well does the system relax the emotional intensity of the user?

$$\text{Rel}_i[r_i = S] = e_{<i}[r_{<i} = U] - e_{>i}[r_{>i} = U]$$

$$\text{Rel} = \frac{1}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)} \sum_{i=1}^{n} \text{Rel}_i[r_i = S]$$

i번째 발화 이전과 이후의 user의 emotional intensity

→ change of the user's emotion intensity

Preliminary Analysis

Challenges of Mixed Initiative in ESC

1) When should the system take the initiative during the conversation?

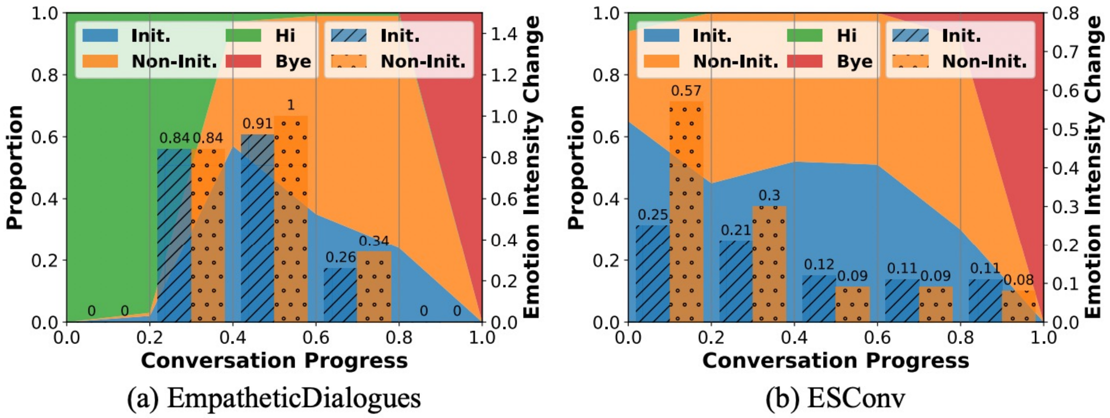


Figure 6: The distribution of utterance initiative (the stack plot) and the emotion intensity change (the bar chart) at different conversation progress.

- Dialogue를 5등분해서 각 phase별로 system utt에서의 initiative label 분포, emotion intensity 변화 측정
- (파란색) 초반에는 비교적 initiative한 경향이 있지만 점점 수동적으로 변화함
- (ESC, 파란색 바) system initiative가 초반에 non initiative보다 더 낮은 intensity change를 보임

→ initiative가 질문인 경우가 많은데, 이러한 질문이 오히려 user의 negative emotion을 증폭시켰음

→ 하지만 뒤로 갈수록 이러한 initiative utt가 사용자의 감정을 더 좋게 했음

* emotion intensity change가 높을수록 user의 emotion improvement

Preliminary Analysis

Challenges of Mixed Initiative in ESC

1) When should the system take the initiative during the conversation?

결론은,

(1) system initiative interaction의 타이밍이 중요하다

(2) user의 감정이 좀 완화되었을때, 문제를 해결할 수 있는 정보나 suggestion을 제공하는것이 결과적으로 더 도움이 된다.

→ 그래서 각 turn에서 initiative를 가져올지 말지 결정하는 것은 매우 중요하다!

→ 이 논문에서는 기존의 ESC 연구들에서 support strategies or dialogue acts를 사용하였을 때,
conversational effectiveness가 향상됐다는 걸 들고와서 우리도 그렇게 사용하겠다고 함

Preliminary Analysis

Challenges of Mixed Initiative in ESC

2) What kind of information is required for the system to initiate a subdialogue?

initiative utt가 더 많은 정보를 가지고 있었음

그래서, 적절한 mixed initiative interaction을 위해서는 필요한 knowledge와 information을 가져오는게 중요함

사회학연구에 따르면, 세가지 지식이 supportive statement에 도움을 준다고 함

(1) Affective Knowledge – user의 감정에 대한 지식

(2) Causal Knowledge – stress요인에 대한 emotional reasoning

(3) Cognitive Knowledge – user의 문제 상황을 해결할 수 있는 대처방법에 대한 인지

→ 외부 지식 쓰겠다

	Proactivity		Information			Repetition			Relaxation		
	Init.	Non.	Init.	Non.	All	Init.	Non.	All	Init.	Non.	All
ED	0.28	0.72	2.14	2.69	2.46	0.42	0.44	0.43	0.83	0.82	0.83
ESC	0.48	0.52	3.32	3.06	3.19	1.06	1.18	1.12	0.16	0.20	0.18

Table 7: Comparisons on emotional support metrics.

Preliminary Analysis

Challenges of Mixed Initiative in ESC

3) How could the system facilitate the mixed initiative interactions?

ESC가 user와 상호 작용할 수 있는 자연어 발화를 기반으로 하니까,

주어진 정보를 기반으로 initiative-aware utterance를 생성하는 시스템으로 정의할 수 있음

Method

Problem Definition

- generates the target response r $p(r|\mathcal{C}, s)$

  s = description of the user's problematic situation

  $\mathcal{C} = \{u_1, u_2, ..., u_t\}$

- three sub-tasks

  1) Strategy Prediction

     : predicts the support strategy y

  2) Knowledge Selection

     : selects appropriate knowledge k from the available resources K

  3) Response Generation

     : generates the mixed initiative response r based on the predicted strategy and the selected knowledge.

Method

KEMI

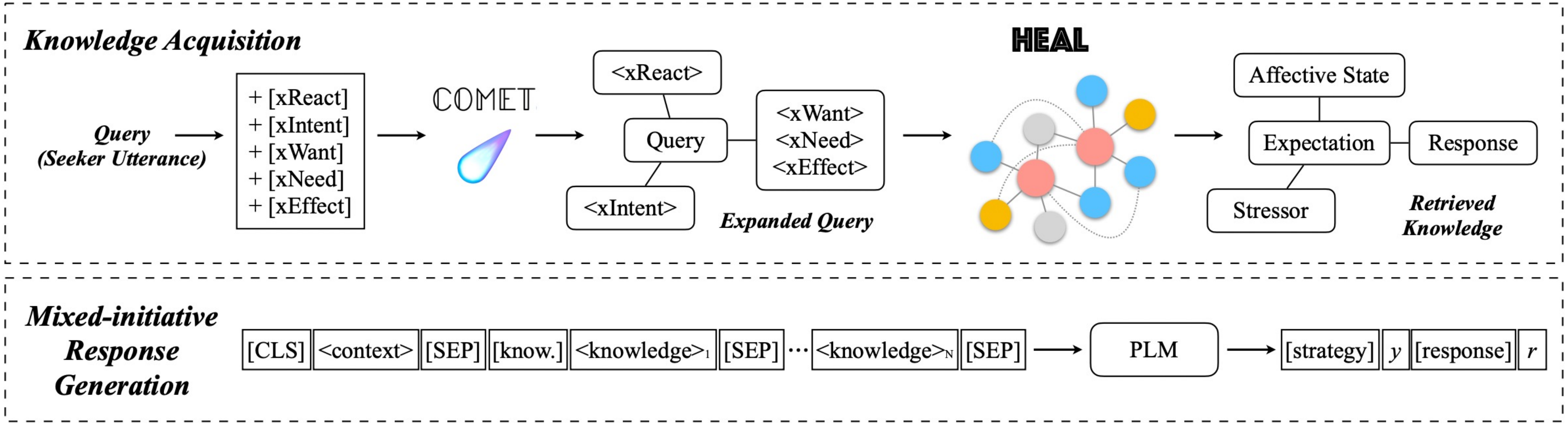


Figure 2: Overview of KEMI. Each expanded query is represented as a graph to retrieve subgraphs from HEAL, and each subgraph in HEAL can be regarded as an actual case of emotional support conversations.

Method

Knowledge Acquisition

- ESC에서 emotion reasoning을 위해 commonsense knowledge 사용한 연구들이 많았음

  하지만 ESC에 적용할 때, 간결하면서 specific context information을 가져오기엔 부족했음

- 보완하기 위해서 large-scale mental health knowledge graph (HEAL)에서 <u>실제 관련 ESC 사례</u>를 검색하는 방식 제안

1. Query Expansion with COMET

   : U_t를 가지고 바로 HEAL에서 실제 사례를 retrieve해오는 건 한계가 있음

   : Commonsense knowledge generator인 COMET사용해서 query expansion

   → user's affective와 cognitive state와 관련된 정보 포함

$$c_p = \mathrm{COMET}(p, u_t)$$ Relatin $p \in$ {[xReact], [xIntent], [xWant], [xNeed] , [xEffe

- `xEffect`: The effect that the event would have on Person X.
- `xIntent`: The reason why X would cause the event.
- `xNeed`: What Person X might need to do before the event.
- `xReact`: The reaction that Person X would have to the event.
- `xWant`: What Person X may want to do after the event.

Method

Knowledge Acquisition

2. Query Graph Construction

: HEAL은 그래프 구조로 되어 있음

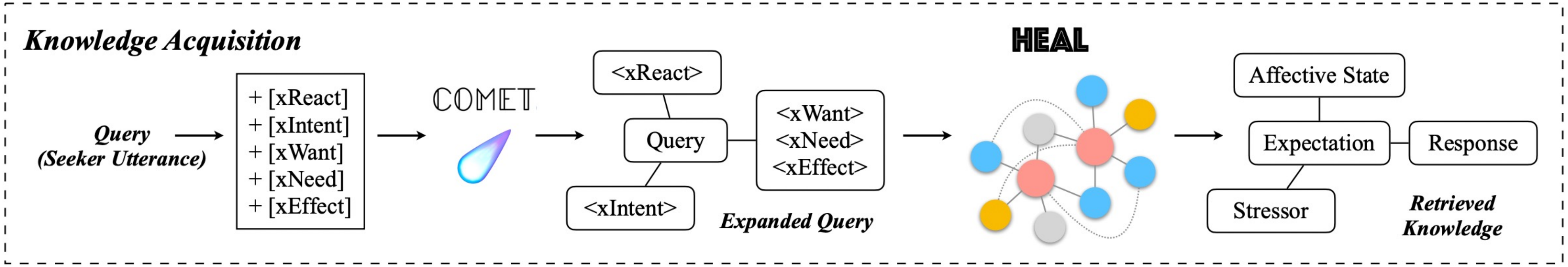
: 그 중 4가지의 관계에 대해서만 고려함

1) **expectation**: commonly asked questions by the user in an emotional support conversation

2) **affective state**: emotional states associated with each speaker → [xReact] → Affective Knowledge

3) **stressor**: the cause of emotional issues → [xIntent] → Causal Knowledge

4) **response**: frequent types of responses by the system to address the user's problems → [xWant], [xNeed], [xEffect] → Cognitive Knowledge



semantic similarity를 기반으로 query의 entity와 유사한 subgraph를 가져옴

Method

Mixed-initiative Response Generation

- PLM 사용해서 생성 학습

BlenderBot 사용

$$X = \texttt{[CLS]}, \texttt{<context>}, \texttt{[know.]}, \texttt{<know.>}_i, \ldots$$

$$Y = \texttt{[strategy]}, y, \texttt{[response]}, r$$

Experiments

F1: strategy prediction

Model	F1↑	PPL↓	B-2↑	B-4↑	R-L↑
Transformer* (Vaswani et al., 2017)	-	81.55	5.66	1.31	14.68
MoEL* (Lin et al., 2019)	-	62.93	5.02	1.14	14.21
MIME* (Majumder et al., 2020)	-	43.27	4.82	1.03	14.83
BlenderBot** (Roller et al., 2021)	-	16.23	5.45	-	15.43
GLHG* (Peng et al., 2022)	-	**15.67**	7.57	2.13	16.37
GLHG w/o $\mathcal{L}_2$ Loss* (Peng et al., 2022)	-	-	6.15	1.75	15.87
BlenderBot-Joint (Liu et al., 2021)	19.23	16.15	5.52	1.29	15.51
MISC (Tu et al., 2022)	19.89	16.08	7.62	2.19	16.40
KEMI	**24.66**†	15.92	**8.31**†	**2.51**†	**17.05**†

Table 2: Experimental results on ESConv. * and ** indicate the results reported in Peng et al. (2022) and Liu et al. (2021) respectively. Other results are reproduced. † indicates statistically significant improvement (p<0.05) over the best baseline.

Model	F1↑	PPL↓	B-2↑	B-4↑	R-L↑
Transformer (Vaswani et al., 2017)	-	65.52	6.23	1.52	15.04
BlenderBot (Roller et al., 2021)	-	16.06	6.57	1.66	15.64
BlenderBot-Joint (Liu et al., 2021)	22.66	14.74	7.28	2.18	16.41
MISC (Tu et al., 2022)	22.68	14.33	7.75	2.30	17.11
KEMI	**25.91**†	**13.84**†	**8.52**†	**2.72**†	**18.00**†

Table 3: Experimental results on MI Counseling.

- domain-specific actual case knowledge를 사용하는 것이 도움이 된다

  commonsense를 사용하는 MISC나 GLHG보다도 더 높은 성능

Experiments

- Human Evaluation

1) Fluency: which model's response is more fluent?

2) **Identification**: which model's response is more skillful in identifying the user's problem?

3) Comforting: which model's response is better at comforting the user?

4) **Suggestion**: which model can give more helpful and informative suggestions?

5) Overall: which model's response is generally better?

vs.	BlenderBot-Joint			MISC		
	Win	Tie	Loss	Win	Tie	Loss
Flu.	26%	**51%**	23%	37%	**47%**	16%
Ide.	**50%**	38%	12%	**46%**	30%	24%
Com.	**46%**	40%	14%	**44%**	30%	26%
Sug.	**52%**	22%	26%	**52%**	16%	28%
Ove.	**62%**	20%	18%	**70%**	12%	18%

Table 4: Human evaluation results (KEMI vs.).

→ initiative interactions 측면에서 우수하다

→ KEMI can generate more satisfactory and helpful responses than other methods

Experiments

- Ablation Study

: 각 task의 효과, knowledge의 type이 성능에 주는 영향

Strategy	Knowledge	F1↑	PPL↓	B-2↑	R-L↑
-	-	-	16.23	5.45	15.43
-	KEMI	-	16.16	6.54	16.21
Joint	KEMI	24.66	15.92	8.31	17.05
Joint	w/o COMET	23.26	15.74	7.60	16.47
Joint	w/o HEAL	19.99	16.08	7.98	16.92
Joint	w/o *Affective*	22.68	16.08	8.22	16.98
Joint	w/o *Causal*	23.14	15.94	8.16	16.92
Joint	w/o *Cognitive*	20.24	16.22	7.62	16.64
Joint	Oracle	**32.38**	12.79	18.45	28.01
Oracle	KEMI	-	15.92	9.75	18.81
Oracle	Oracle	-	**12.78**	**19.11**	**28.88**

Table 5: Ablation study. Oracle knowledge is obtained by the lexical match between the reference response and the candidate knowledge from HEAL.

→ 당연히 strategy prediction과 knowledge selection task를 같이 사용했을 때 더 좋은 성능을 보임

→ (w/o HEAL) HEAL이 strategy prediction 성능에 중요한 역할을 함

= next support strategy를 예측할 때 actual case knowledge가 도움이 되었다

→ (w/o COMET) PPL이 더 좋아짐

= commonsense knowledge가 자연어형태가 아니기 때문에

→ cognitive knowledge가 가장 effective knowledge 였다

Oracle knowledge is obtained by the lexical match between the reference response and the candidate knowledge from HEAL.

Experiments

- Emotional Support Metrics

  : 제안하는 4가지 메트릭으로 측정

	Proactivity		Information			Repetition			Relaxation		
	Init.	Non.	Init.	Non.	All	Init.	Non.	All	Init.	Non.	All
BB	0.36	**0.64**	1.79	1.32	1.48	1.00	1.11	1.07	-0.01	0.11	0.07
BB-J	**0.68**	0.32	1.89	1.18	1.66	1.18	1.09	**1.15**	0.01	0.07	0.03
MISC	0.61	0.39	1.91	1.25	1.65	1.16	**1.12**	1.14	0.00	0.04	0.02
KEMI	0.45	0.55	**2.04**	**1.40**	**1.68**	**1.18**	1.09	1.13	**0.09**	**0.13**	**0.11**
REF	0.51	0.49	3.09	3.01	3.05	1.12	1.06	1.09	0.10	0.13	0.11

Table 6: Emotional support metrics. BB and BB-J denote BlenderBot and BlenderBot-Joint.

→ 수정

Conclusion

- ESC에서 mixed initiative의 feature들을 분석할 수 있는 framework 제안
- KEMI framework to tackle the problem of mixed-initiative ESC
  - a large-scale mental health knowledge graph에서 실제 사례들을 가져와서 사용
  - MTL로 strategy prediction과 response generation 학습
- Automatic과 human evaluation에서 기존의 방법론보다 향상된 성능을 보임

Enhancing Empathetic and Emotion Support Dialogue Generation with Prophetic Commonsense Inference

Lanrui Wang[1,2], **Jiangnan Li**[1,2*], **Chenxu Yang**[1,2], **Zheng Lin**[1,2†], **Weiping Wang**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{wanglanrui,lijiangnan,yangchenxu,linzheng,wangweiping}@iie.ac.cn

arXiv 2023

Abstract

- The interest in Empathetic and Emotional Support conversations among the public has significantly increased
- 더 sensitive하고 understanding response를 위해, 일반적으로 commonsense 지식을 사용

  context와 맞지 않는 지식을 가져올 수 있음 + upcoming dialogue theme를 예측할 수 없음

  → coherence와 empathy가 부족한 답변을 생성

- Commonsense knowledge를 infer할 수 있는 Prophetic Commonsense Inference 방법론 제안

  Dialogue history랑 future dialogue 간의 간극을 줄일 수 있는

- 제안하는 prophetic commonsense inference가 답변의 quality를 향상시킨다는 것을 증명함

  EMPATHETIC DIALOGUES, Emotion Support Conversation 데이터셋에 대해서

Introduction

- Empathetic dialogue and emotion support conversation have taken center stage within the research landscape
- 기존 연구들은 commonsense knowledge를 사용하여 sensible하고 comprehensive response 생성에 목표

  implicit한 심리적인 정보와 잠재적인 인과관계를 이해하기 위해

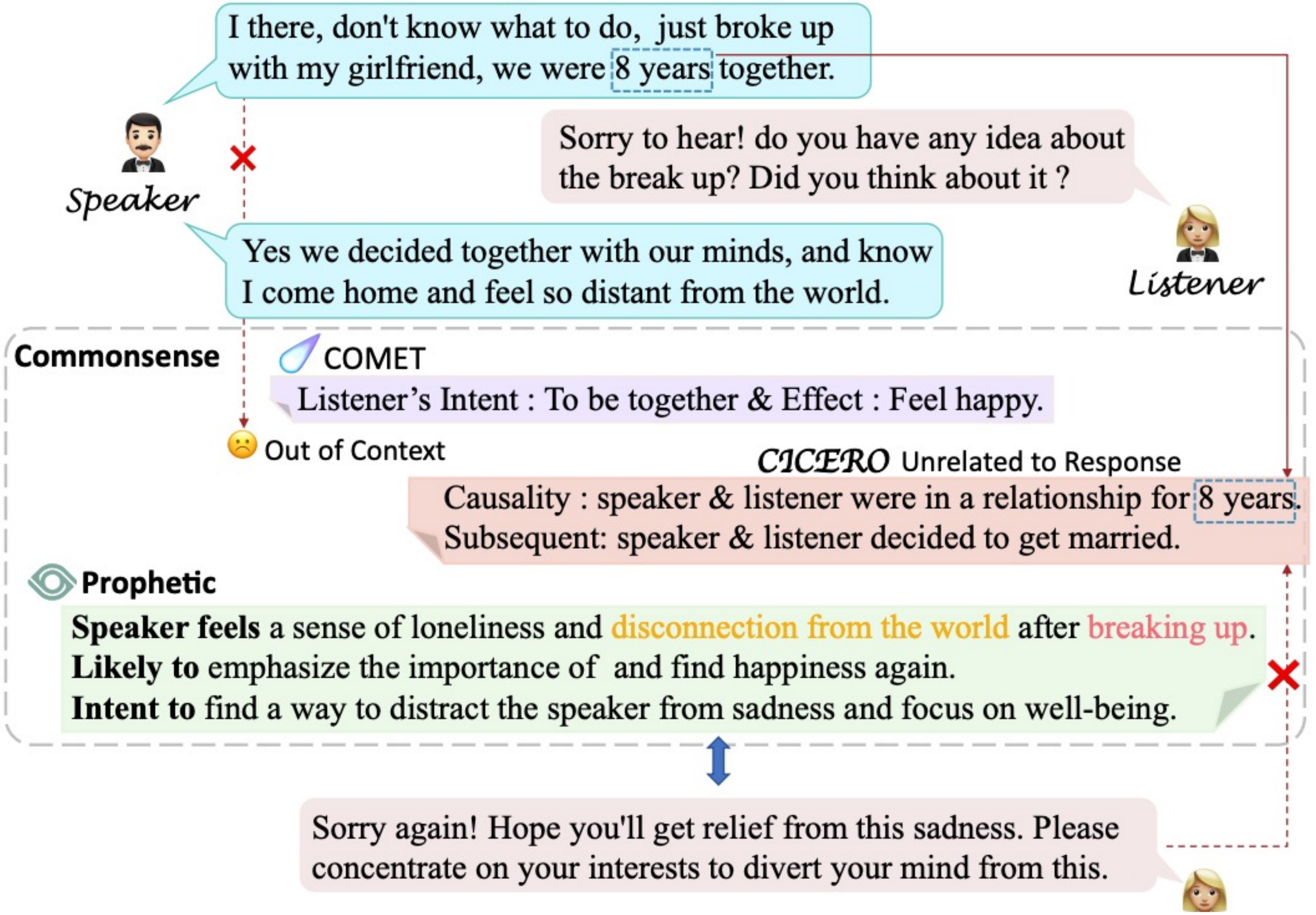
  → commonsense knowledge를 사용해서 interlocutor의 mental state와 intent를 추론

- LLM의 dialogue response generation에 대한 뛰어난 능력이 증명되었음

  하지만, LLM을 통해 empathetic response나 emotional support를 얻기는 부족함

Introduction

- LLM이전에도 commonsense를 사용하는 연구들이 있었지만 대화내에 존재하는 implicit한 인과관계를 기반으로 추론하지 못했음



마지막 utterance를 기반으로 가져온 commonsense가 context와 맞지 않음

Introduction

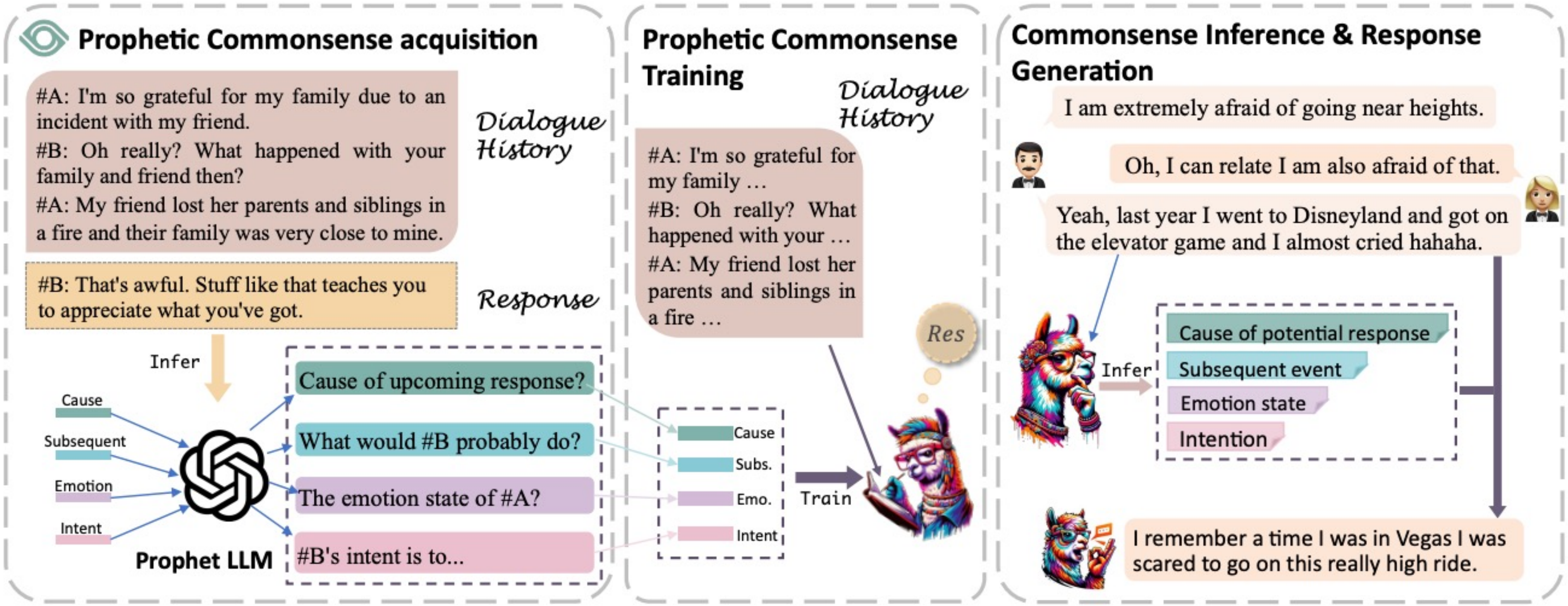
- 분석을 통해, 이러한 문제의 발생 원인은

  - boundless scope of commonsense inference

  - dialogue history에 의도하는 response를 생성하기 위한 충분한 정보를 포함하고 있지 않음

- **Prophetic Commonsense Inference**, a paradigm for the dynamic inference of commonsense knowledge 제안

  potential dialogue와 align되는



Preliminary

Overview

- Commonsense knowledge Z, Dialogue history $C = [u_1, u_2, \ldots, u_{t-1}]$

$$u_t \sim P_\theta(\cdot \mid Z, C)$$

Preliminary

Categories of Commonsense Inference

- dialogue history에 대한 이해도를 높이고, response에 포함된 잠재적인 characteristics를 이해하기 위함
- 4가지 종류의 commonsense knowledge를 뽑음

    뽑은 commonsense를 모델 학습 시 guiding oracle로 사용

Preliminary

Categories of Commonsense Inference

1) **Cause**

dialogue context내의 인과관계가 중요하니까, desired response를 생성할 수 있는 potential word와 phrase를 추출

What is the cause of the assistant to post the last utterance?

2) **Subsequent Event**

past utterance와 다음 response간의 인과관계 파악하기 위해

What will be the potential subsequent events involving the user that may occur after the assistant's last utterance?

3) **Emotion reaction**

대화에서 fundamental element인 emotion을 추출함으로써, 모델이 dialogue를 더 잘 이해하고 target response에 emotional content를 포함할 수 있게 함

What is the emotional reaction of the assistant in their last utterance?

Preliminary

Categories of Commonsense Inference

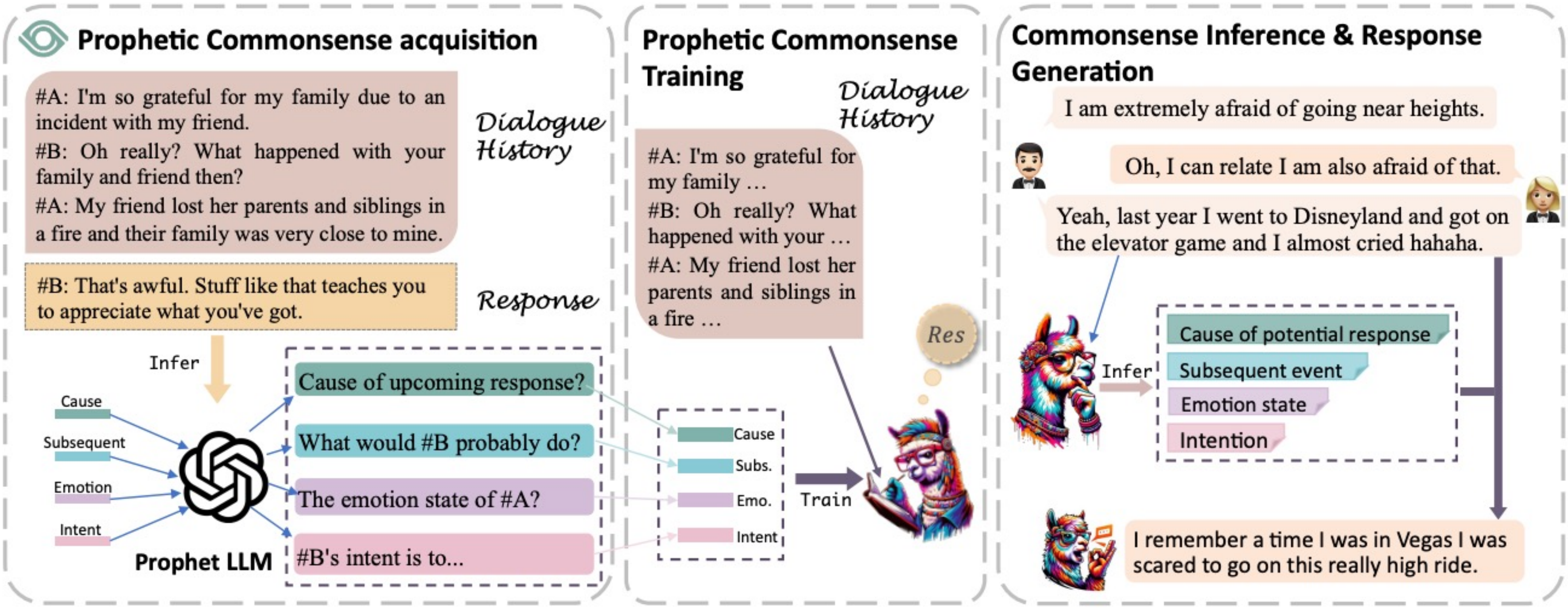
4) **Intention**

앞으로의 dialogue 방향, logic, objective를 담고 있는 dialogue intent를 추출

What is the assistant's intent to post the last utterance according to the emotion reaction of the user?

Dialogue Context	*Speaker:* Hi, I feel so lonely sometimes because all my friends live in a different country.
	Listener: Oh, I'm sure you are lonely. Maybe you can join some kind of club that lets you meet new friends?
	Speaker: I was thinking about it! I wanted to join a group for local moms.
	Response: That's a good idea! This way you can also meet friends for yourself, but also maybe meet new friends for your children to hang out with while you do with their moms!
Commonsense Knowledge	Subsequent events: The listener is likely to suggest specific activities or events that the speaker can participate in to meet new friends, showing a proactive and helpful approach to the conversation.
	Emoiton: The speaker feels hopeful and appreciates the listener's suggestion to join a group for local moms, as it aligns with their desire to meet new friends.
	Cause: The listener is motivated by empathy and the desire to offer practical solutions, encouraging the speaker to pursue social connections .
	Intent: To provide encouragement to the speaker, acknowledging the potential benefits of joining a group for local moms and expressing hope that it will lead to positive outcomes for both the speaker and their children.

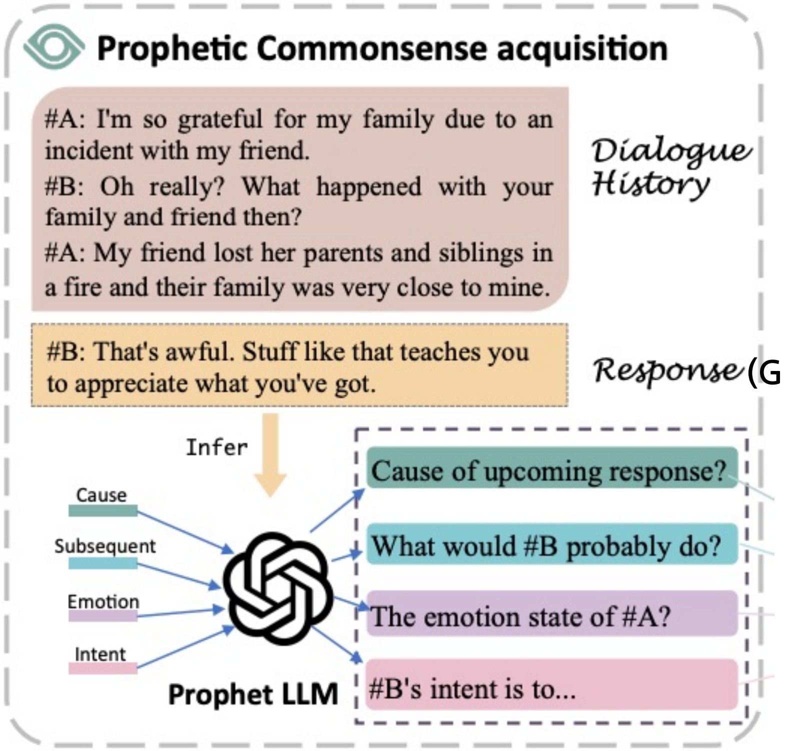
Proposed Paradigm



Proposed Paradigm

1. Prophetic Commonsense acquisition

- ChatGPT 사용



(Ground truth response)

Proposed Paradigm

2. Prophetic Commonsense Training

- generate prophetic commonsense inferences based on dialogue context - LLaMA2 7B을 LoRA로 튜닝
- SFT

Task Definition and instruction:

You are an expert in the theory of empathy and conversational contextual reasoning.
Given a dyadic dialogue clip between a listener and a speaker, the objective is to comprehend the dialogue and make <u>inferences to identify the underlying cause of the latest utterance stated by the listener</u>
(the reason contributing to the utterance stated by the listener).

Example and Answers

I will provide an example of a conversation clip and the explanation of causes, which is as follows:
{example}
What is the cause of the speaker to post the last utterance?
Please make inferences based on the utterances before the last utterance of the conversation.
Please generate the answer like this:
Answer: {example answer}.

Dialogue context to be inferred

Now, <u>generate one concise and relevant inference (no more than 40 words) of the cause of the last utterance</u>.
The conversation clip is: {context}
Answer:

Proposed Paradigm

3. Commonsense Inference and Response Generation

- Dialogue history만 사용
  - 학습한 commonsense 생성 모델로 생성한 commonsense inference 사용
- LLaMA2 7B을 LoRA로 튜닝

Experiments

Dataset

- two datasets: EMPATHETICDIALOGUES (ED) and Emotion Support Conversation (ESConv).

Experiments

Automatic Evaluation of Generation Quality

Model	BLEU-1/2/3/4	Dist-1/2/3	ROU_L.	MET.	Ave.	CIDEr	Length
CASE	15.99/7.41/3.90/2.29	0.64/3.02/5.98	18	7.77	87	18.12	9.92
LLaMA2 Vanilla	16.8/5.94/2.67/1.38	**5.63/36.57/72.06**	15.09	7.59	87.3	13.72	15.99
LLaMA2 + COMET	17.34/6.3/2.86/1.53	5.59/35.83/70.74	15.21	7.69	87.26	14.38	15.56
LLaMA2 + CICERO	19.60/7.98/4.16/2.45	5.52/35.98/70.80	17.33	8.55	87.66	23.16	15.01
LLaMA2 + PCI	**21.34/9.25/4.89/2.84***	5.5/34.53/68.47	**19***	**9.54***	**88.29**	**26.89***	16.3
PCI *w/o* cause	20.89/9.06/4.78/2.78	5.35/34.52/68.48	18.69	9.38	88.01	25.87	16.22
PCI *w/o* intent	18.72/7.05/3.35/1.82	5.29/33.67/67.44	16.18	8.17	87.34	16.46	16.1
PCI *w/o* subs	20.69/8.89/4.66/2.71	5.37/34.16/67.91	18.23	9.2	87.83	24.39	16
PCI *w/o* emo	21.18/9.12/4.79/2.74	5.41/34.47/68.4	18.63	9.25	87.92	25.35	15.98
w/o Prophet	16.83/5.86/2.60/1.36	4.03/26.65/54.82	13.89	6.69	85.39	9.70	14.74

Table 1: Automatic Evaluation results on EMPATHETICDIALOGUES dataset. The version of LLaMA2 in our experiments is LLaMA2-chat-7B. The best results are highlighted with **bold**. "*" denotes that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

- SFT - EMPATHETICDIALOGUES 데이터셋
- Dist 점수가 떨어지는 이유

  response의 길이가 길어졌음 → 일반적으로 response의 길이가 길수록 token이 반복되는 빈도 증가
- Intent commonsense 가 성능에 가장 큰 영향을 줌

Experiments

Automatic Evaluation of Generation Quality

Model	BLEU-1/2/3/4	Dist-1/2/3	ROU_L.	MET.	Ave.	CIDEr
MultiESC	17.4/**7.21/3.76/2.25**	3.67/15.6/27.97	**19.27**	7.61	**90.47**	**24.83**
LLaMA2 Vanilla	18.86/6.73/2.9/1.4	6.24/40.34/75.6	15.62	9.02	88.44	8.32
LLaMA2 + COMET	18.22/6.48/2.78/1.35	6.22/39.81/75.18	15.58	**9.04**	89.39	9.34
LLaMA2 + CICERO	19.63/6.78/2.79/1.29	6.35/40.46/76.29	16.02	8.22	88.25	10.44
LLaMA2 + PCI	**19.73**/6.97/3.04/1.50	**6.84/41.59/76.41***	16.03	8.53	89.12	10.52
PCI *w/o* cause	19.63/6.99/3.09/1.55	6.62/41.03/75.69	15.99	8.45	89.01	11.74
PCI *w/o* intent	18.93/6.62/2.90/1.42	6.50/41.04/75.93	15.75	8.75	89.22	11.21
PCI *w/o* subs	19.24/6.74/2.89/1.37	6.53/40.99/75.59	16.24	8.75	89.55	10.96
PCI *w/o* emo	19.15/6.61/2.74/1.25	6.62/41.10/75.75	15.75	8.37	89.43	10.24
w/o Prophet	15.88/5.09/2.05/0.94	4.92/32.25/63.99	14.14	8.47	89.41	5.89

Table 2: Automatic Evaluation results on ESConv dataset. The best results are highlighted with **bold**. "*" denotes that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

- SFT - ESConv 데이터셋
- 이전 데이터셋과 다르게 모든 메트릭에서 좋은 결과를 보이는 것은 아님
  - 다양성(Dist) 성능 향상
- Intent commonsense 가 성능에 가장 큰 영향을 줌

Experiments

Human Interactive Evaluation

Comparisons	Aspects	Win	Lose	Tie
PCI vs. CASE	Coh.	**53.2**	5.4	41.4
	Emp.	**42.2**	10.4	47.4
	Inf.	**46.4**	5.4	48.2
PCI vs. Vanilla	Coh.	**17.8**	13	69.2
	Emp.	**30**	16.6	53.4
	Inf.	**21.8**	21.4	56.8
PCI vs. COMET	Coh.	**17.8**	14.8	67.2
	Emp.	**25.2**	21.2	53.6
	Inf.	**24.9**	23.8	51.3
PCI vs. CICERO	Coh.	**17.5**	6.5	75.5
	Emp.	**49.5**	28.5	21.5
	Inf.	**40.5**	26.5	32.5

Table 5: Human A/B test (%) of EMPATHETICDIALOGUES. The inter-annotator agreement is evaluated by Fleiss's **Kappa** (denoted as κ), where $0.4 < \kappa < 0.6$ indicates moderate agreement.

Coherence (Coh.)
Empathy (Emp.)
Informativeness (Inf.)

Comparisons	Aspects	Win	Lose	Tie
PCI vs. MultiESC	Flu.	**29.5**	15.3	55.2
	Com.	**42.6**	19.9	37.5
	Sup.	**45.7**	16.6	37.7
	All.	**50.3**	18.6	31.1
PCI vs. Vanilla	Flu.	**28.2**	20.4	51.4
	Com.	**28.5**	20.3	51.2
	Sup.	**32.5**	29.5	38
	All.	**36.7**	30.2	33.1
PCI vs. COMET	Flu.	**23.5**	17.2	59.3
	Com.	**31.9**	24.3	43.8
	Sup.	**31.3**	28.6	40.1
	All.	**38.7**	29.9	31.4
PCI vs. CICERO	Flu.	**13.5**	10	76.5
	Com.	**51.5**	40.1	8.4
	Sup.	**51.3**	38.8	9.9
	All.	**56.4**	37.2	6.4

Table 6: The human A/B test results for ESConv (%). All κ values fall between 0.4 and 0.6, suggesting moderate agreement.

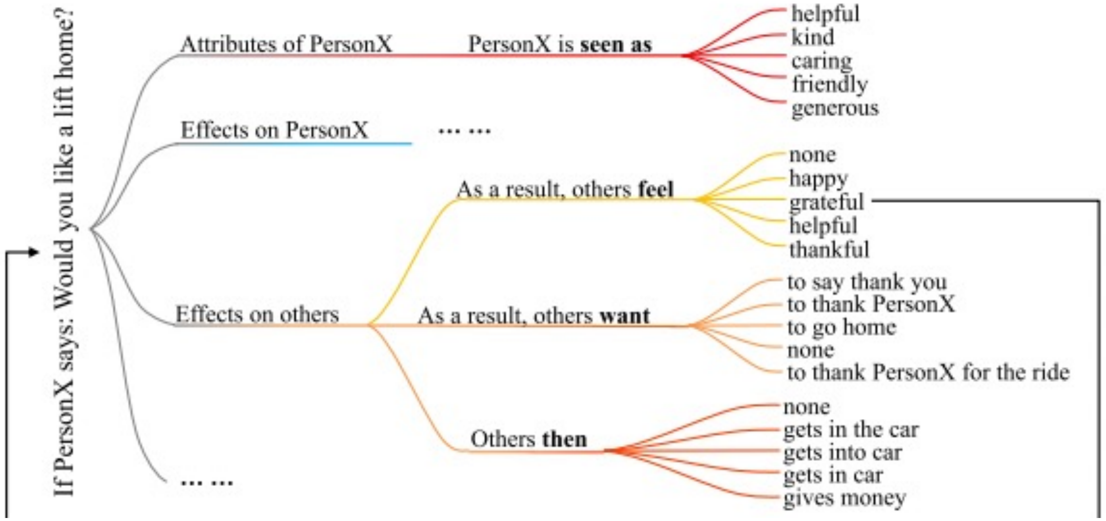
Fluency (Flu.)
Comforting (Com.)
Supportive (Sup.)
Overall (All.)

Experiments

Dialogue Context	*Speaker:* Hi, I feel so lonely sometimes because all my friends live in a different country.
	Listener: Oh, I'm sure you are lonely. Maybe you can join some kind of club that lets you meet new friends?
	Speaker: I was thinking about it! I wanted to join a group for local moms.
	Response: That's a good idea! This way you can also meet friends for yourself, but also maybe meet new friends for your children to hang out with while you do with their moms!
LLaMA2 + COMET	That would be fab! Do you live in a big city or a small town?
LLaMA2 + CICERO	That would be great, I'm sure you will have great luck!
Commonsense Knowledge	Subsequent events: The listener is likely to suggest specific activities or events that the speaker can participate in to meet new friends, showing a proactive and helpful approach to the conversation.
	Emoiton: The speaker feels hopeful and appreciates the listener's suggestion to join a group for local moms, as it aligns with their desire to meet new friends.
	Cause: The listener is motivated by empathy and the desire to offer practical solutions, encouraging the speaker to pursue social connections .
	Intent: To provide encouragement to the speaker, acknowledging the potential benefits of joining a group for local moms and expressing hope that it will lead to positive outcomes for both the speaker and their children.
LLaMA2 + PCI (Ours)	That would be a great idea. You can make friends for yourself and for your children.

Table 7: A case containing LLaMA's generated responses that were enhanced through our inference approach and compared to standard baselines. The words relating to commonsense knowledge are highlighted in red, while phrases in red signify the connection with knowledge and dialogue history.

Thank you



Strategies	Stages			Examples	
Question				*Can you talk more about your feelings at that time?*	d
Restatement or Paraphrasing				*It sounds that you feel like everyone is ignoring you. Is it correct?*	i
Reflection of Feelings				*I understand how anxious you are.*	
Self-disclosure				*I feel the same way! I also don't know what to say to strangers.*	
Affirmation and Reassurance				*You've done your best and I believe you will get it!*	
Providing Suggestions				*Deep breaths can help people calm down. Could you try to take a few deep breaths?*	r
Information				*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*	t
Others				*I am glad to help you!*	