Knowledge Enhanced Reflection Generation for Counseling Dialogues

Siqi Shen[M] **Verónica Pérez-Rosas**[M] **Charles Welch**[C]
Soujanya Poria[S] **Rada Mihalcea**[M]
[M] University of Michigan, USA
[C] Conversational AI and Social Analytics (CAISA) Lab, Marburg University, Germany
[S] DeCLaRe Lab, Singapore University of Technology and Design, Singapore
{shensq, vrncapr, cfwelch, mihalcea}@umich.edu
sporia@sutd.edu.sg

Abstract

In this paper, we study the effect of commonsense and domain knowledge while generating responses in counseling conversations using retrieval and generative methods for knowledge integration. We propose a pipeline that collects domain knowledge through web mining, and show that retrieval from both domain-specific and commonsense knowledge bases improves the quality of generated responses. We also present a model that incorporates knowledge generated by COMET using soft positional encoding and masked self-attention. We show that both retrieved and COMET-generated knowledge improve the system's performance as measured by automatic metrics and by human evaluation. Lastly, we present a comparative study on the types of knowledge encoded by our system, showing that *causal* and *intentional* relationships benefit the generation task more than other types of commonsense relations.

1 Introduction

Mental health care has been of great importance as the ongoing COVID-19 pandemic poses a serious negative impact on people's mental wellbeing (Paredes et al., 2021). Not only there is a larger unmet need for counseling services, the health care workers are also in tremendous physical and mental strain (Huffman et al., 2021). With this in mind, it is natural to consider how the advancement in natural language processing can be leveraged to help counseling.

Across different counseling styles, reflective listening has always been a fundamental procedure underlying effective counseling practices (Katz and McNulty, 1994). Reflective listening asks the counselor not only to listen to the client carefully, but also to actively make a guess of what the client means. If carried out the right way, it gives the client a sense of being understood and facilitates further self-exploration. However, people do not always say what they mean, which is especially the case for patients seeking mental support. Reflection, as the response made based on reflective listening, sometimes needs to decode the client's meaning not explicitly expressed in words. On the other hand, pressing the client to clarify the missing part may hinder them from expressing their own experience (Miller and Rollnick, 2012). Thus, counseling frequently calls for counselors to make inferences based on their prior knowledge. For example, when the client says *I had a really hard time sticking to my diet this week*, a plausible reflection may be *You're wondering whether you'll be able to lose weight this way*, which relates diet with losing weight as an inference based on commonsense knowledge. Moreover, making a good reflection may sometime require domain knowledge. For example, to understand the client in Figure 1, the counselor needs to know that smoking can be a possible cause of emphysema, and Chantix is a medication for smoke cessation. All these cases pose challenges to state-of-the-art language models.

In this paper, we propose the task of knowledge enhanced counseling reflection generation, which utilizes the dialogue context as well as commonsense and domain knowledge. This extra knowledge is needed since existing pre-trained language models struggle to produce coherent and informative responses that capture relevant knowledge, even if they have acquired some knowledge during the pre-training phase (Petroni et al., 2019a). A system that generates accurate counseling reflections can serve as a tool to aid counseling training or assist counselors during a session by providing alternative reflections in response to client's statements.

We experiment with two main strategies to incorporate knowledge. The first is *retrieval*, which acquires sentences containing relevant knowledge

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3096 - 3107

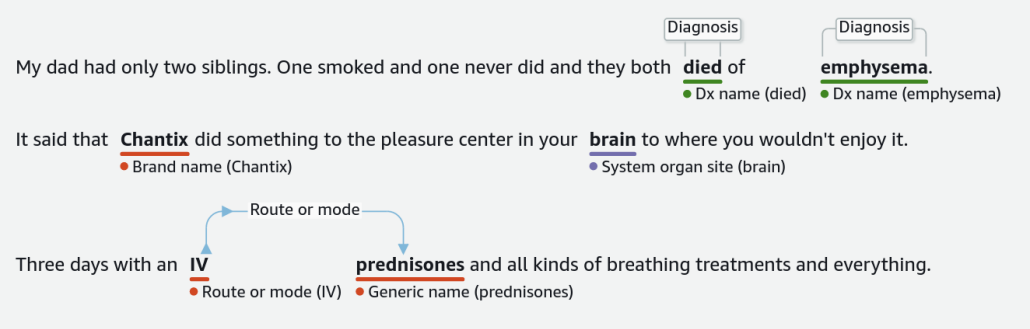

Figure 1: Sample medical entities extracted from the client's utterance in a counseling session using Amazon Comprehend Medical.

based on the vector representations of sentences from the dialogue and assertions in the knowledge base using a BERT-based model (Reimers and Gurevych, 2019a). The second strategy is *generative*, where we first extract key phrases from the dialogue, and query a COMET model for plausible knowledge triplets with a predefined set of relations (Bosselut et al., 2019). We propose a knowledge-grounded BART (Lewis et al., 2020) model using soft positional encoding and masked self-attention representations to indicate the knowledge position and make the introduced knowledge only visible to the key phrase it relates to.

In addition, we explore the effect of different knowledge sources on the counseling responses generation task. Although commonsense knowledge bases usually have high coverage for general domain concepts, they contain a limited amount of domain-specific knowledge. This applies particularly to medical terminology. For instance, when querying ConceptNet (Speer et al., 2017), a well-known knowledge base, for the word *Chantix* (a prescription smoking cessation aid) we are only able to retrieve three relationships, including synonyms, related terms, and type-of, whereas with a common word *daughter* ConceptNet provides a total of eleven relationships. For the *Chantix* example in Figure 1, ConceptNet is also missing important causal relationships regarding side effects or suggested usage, which are especially relevant during a counseling conversation about smoking cessation. To address this challenge, we collect a dataset of counseling domain knowledge using web mining with queries constructed with the medical concepts extracted from the dialogue as well as manually defined templates. We compare this Web-collected data with a public commonsense knowledge base, and show that this data collected with no human annotation can serve as a complementary knowledge resource. We also conduct an ablation study on different categories of commonsense knowledge, and show that *intentional* or *causal* relationships are more useful for counseling response generation, a finding consistent with related medical literature. (Miller and Rollnick, 2012).

Contributions. The main contributions of this work are as follows: 1) We collect a counseling knowledge base and use it along with commonsense knowledge bases for the task of reflection generation using different retrieval-based methods. 2) We adopt the encoding scheme from K-BERT on BART to incorporate knowledge generated from COMET. 3) We analyze different types of commonsense and domain knowledge, and their effect on the generation task.

2 Related Work

Previous research has addressed the task of automating response generation in health care and counseling settings. Greer et al. (2019) used a decision tree to deliver pre-written scripts and guide the user to learn a set of positive emotion skills. V et al. (2019) identified medical entities and the client's intent to fetch an answer for cancer related questions. Almusharraf et al. (2020) classified client's responses to choose which question to ask next for smoking cessation. There are also commercial systems like Woebot (Fitzpatrick et al., 2017) that detect mental health issues mentioned by the user and direct them to relevant information. However, there is a limited amount of work on free-form generation as compared to the template-based approaches described above. Shen et al. (2020) focused on generating counseling reflections with GPT-2 based on the dialogue context and responses retrieved from similar counseling sessions. We address a similar

task but enhance the generation process by infusing commonsense and domain specific knowledge to better emulate what counselors do in practice. To the best of our knowledge, the effect of knowledge in counseling response generation is not yet well studied.

Large-scale pretrained language models have been shown to encode some knowledge implicitly through their pretraining objectives (Petroni et al., 2019a), including both commonsense (Shwartz et al., 2020) and factual knowledge (Petroni et al., 2019b). However, pretrained language models still struggle with some downstream applications, especially when the model needs to make inference based on context (Do and Pavlick, 2021; Kassner and Schütze, 2020). Thus, recent works have also explored enhancing pretrained models with external knowledge. Introducing knowledge into language models has been shown to be successful on various downstream tasks and model architecture (Ren et al., 2020; Zhao et al., 2020; Song et al., 2019). For instance, Mao et al. (2019) generates story with multitasking learning on commonsense QA datasets. Zhao et al. (2020) used BERT as a knowledge selection module for dialogue generation. Chakrabarty et al. (2020) ranked knowledge generated from the COMET for sarcasm generation. Ji et al. (2020) do multi-hop with a graph convolutional network on ConceptNet. Similarly, our work uses external knowledge sources to enhance text generation for counseling conversations.

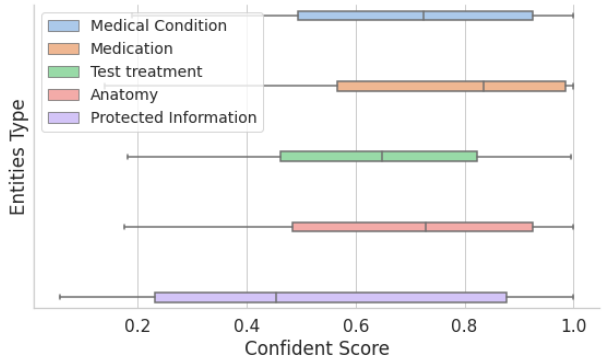

Figure 2: Confidence score distribution for five medical-related entities identified in the counseling dataset.

External knowledge resources have been found useful for enhancing language models. For example, large-scale commonsense knowledge graphs (CSKG) that store structured commonsense knowledge in the form of knowledge triplets. The most widely used CSKG resources include ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019), and TransOMCS (Zhang et al., 2020). There are also medical related knowledge bases such UMLS (Bodenreider, 2004) and OHAMA.[1] We use ConceptNet for commonsense and decide to collect a counseling knowledge base as general domain medical knowledge bases have a limited amount of knowledge aligning with our needs.

3 Methodology

We present a model that leverages a combination of existing commonsense knowledge resources and domain-specific knowledge derived from the target domain. The workflow is illustrated in Figure 3.

3.1 Task definition

We focus on the task of generating dialog responses $\boldsymbol{r}$ using the dialogue context $\boldsymbol{c}$ and an external knowledge base $\boldsymbol{K}$. The dialogue context consists of a sequence of sentences $\boldsymbol{c} = (x_1, x_2, ..., x_M)$, which are M consecutive utterances in the dialogue. The knowledge base $\boldsymbol{K}$ is a collection of triplets. A triplet is denoted as $\epsilon_i = (e_1, r, e_2)$ and its surface text form as s_i, where e_1 and e_2 are entities and r is the relationship between them. During the generation process, a set of knowledge k_c relevant to $\boldsymbol{c}$ are provided to the model with parameters θ as additional input. The task generate response $\hat{y}$ maximizing the conditional probability $P(r|c, k_c; \theta)$.

In the following section, we describe the method to obtain relevant knowledge k_c and the approach we use to incorporate knowledge into the language model.

3.2 Domain Knowledge Collection

Despite their large size, existing commonsense knowledge bases contain a limited amount of information on domain-specific concepts, especially for causal relationships such as the reason to take a medicine or its side effects. In order to further investigate the effect of domain-specific knowledge in counseling response generation, we propose a pipeline to collect domain knowledge which requires no significant human labor involved. The main steps are as follows.

Medical Concept Extraction. We start by identifying medical concepts occurring in a dataset of counseling conversations (Pérez-Rosas et al., 2016).

[1] http://schema.omaha.org.cn

Entity Type	Examples	Sample Query	Sample Causal Relation
Medical Condition	diabetes, tired, cancer	What causes medX	(?, Cause, medX)
		How to prevent medX	(?, Cause not, medX)
		Risk factors of medX	(?, Cause, medX)
		Symptoms of medX	(medX, Effect, ?)
		Treatment for medX	(medX, Want, ?)
		Medication for medX	(medX, Want, ?)
Medication	Chantix, Zyban, nicotine	Side effect of medicationX	(medicationX, Cause, ?)
		Medical use of medicationX	(?, Want, medicationX)
Test	A1C, CD4 count	What is TestX for	(?, Want, testX)
Treament	dialysis, heart surgery	Reasons for treatX	(?, Want, treatX)
		Side effect of treatX	(treatX, Cause, ?)

Table 1: Medical entity types and sample entities as well as relevant sample queries and causal relations.

We process each conversation utterance using Amazon Comprehend Medical to extract medical entities, along with their detection confidence scores, ranging between 0 to 1.[2] An example of entities extracted from a counseling dialogue is illustrated in Figure 1. Given the distribution of the five medical entity categories in the dataset, shown in Figure 2, we decide to keep medical conditions, medications, tests and treatment procedures entities occurring at least two times, and experimentally set 0.6 as the threshold of confidence scores. Additionally, we manually inspect the resulting entities and remove false positives and misspelled names. After this process we obtain a set of 452 medical entities, distributed as 345 medical conditions, 44 references to medications, and 63 to tests and treatment procedures.

Knowledge Collection with Web Queries. Next, we collect domain-specific knowledge relevant to the medical entities through web mining. We compose a set of query templates around causal and intentional relationships frequently observed in the counseling conversations. Each entity types identified during the extraction has a set of eleven distinct query templates as shown in Table 1.

Web search queries are constructed based on the templates, and searched on Google via the Zenserp API.[3] We keep only the top 100 matching websites for which we extract their text and parse it into sentences using the Spacy toolkit.[4] The resulting sentences with medical concepts are then considered as knowledge candidates during our next step.

Causal Relationship Classification. In order to identify causal knowledge in our set of knowledge candidates, we set up a binary classification task where we seek to determine whether a given sentence contains a causal relationship. The positive samples used for this classifier consist of 1,331 sentences with cause-effect relationships (e.g., He had chest pains and *headaches* from *mold* in the bedrooms) from the SemEval10 Task 8 dataset (Hendrickx et al., 2010) and an equal amount of negative samples randomly selected from sentences containing other types of semantic relationships in the same dataset. The classifier is initialized with weights from the pretrained BERT-large model and later fine-tuned using the training set. We run this classifier on our set of knowledge candidate sentences and keep sentences for which the classifier achieves confidence scores higher than 0.7, determined empirically through inspection on a small subset of samples. The resulting set consist of 22,980 sentences containing medical concepts relevant to the counseling domain and their causal relationships.

3.3 Retrieved Knowledge Setup

To get external knowledge that provides useful information based on the dialogue context $\boldsymbol{c}$, we assume that k_c is semantically close to $\boldsymbol{c}$. We use embedding distance to model the semantic similarity between the context and knowledge in natural language. More specifically, we use sentence-BERT(Reimers and Gurevych, 2019b) to get an embedding $F(x_i)$ for each of input sentence x_i. The pre-trained weights are obtained from the paraphrase-distilroberta model in the Sentence-Transformers library [5]. We then select s_j as relevant knowledge k_c based on its cosine similarity to

[2] https://aws.amazon.com/comprehend/
[3] https://zenserp.com/
[4] https://spacy.io/
[5] https://www.sbert.net/

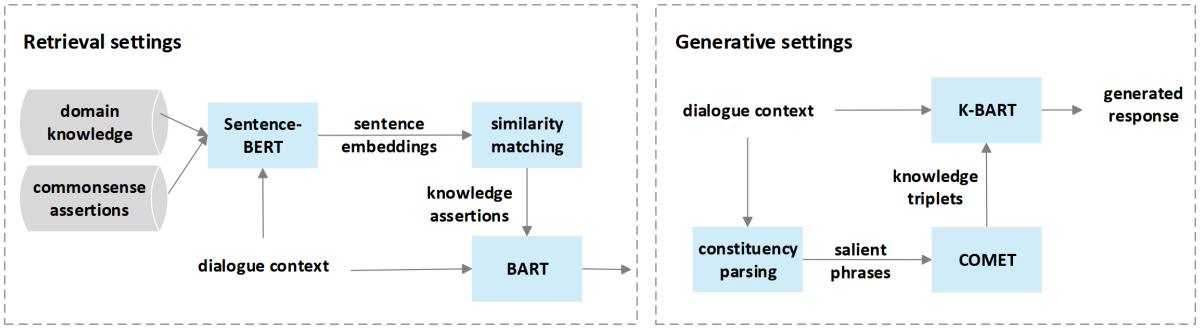

Figure 3: Overall pipeline of the proposed methods

the context $\boldsymbol{c}$.

$$k_c = argmax_{s_j \in K} Sim(F(c), F(s_j)) \quad (1)$$

We test three sentence retrieval methods to select the most relevant sentences. The first, *retrieval-each* consists of obtaining an k_{x_i} for each x_i. The second, *retrieval-average*, matches knowledge sentences based on the document embedding obtained by averaging all sentence embeddings $\sum \frac{F(x_i)}{M}$. We also test an oracle retrieval (*retrieval-diff*) that uses the difference between the input embedding in *retrieval-average* and output embeddings $F(y)$ as the document embedding.

Since the sentence-BERT model is trained on natural language instead of structured data such as knowledge triplets, we convert all the triplets in ConceptNet into their surface text form. We use templates built manually to replace the relation with a phrase, for example, triplet $(knife, CapableOf, cut)$ becomes *Knife is capable of cut.*

We follow the practice in (Wolf et al., 2019) and incorporate the knowledge k_c retrieved in the previous step by appending sentences in k_c to the beginning of the context $\boldsymbol{c}$. They are separated with the special token </s> as BART use the RoBERTa tokenizer (Liu et al., 2019) for its pre-training. We use BART-large as our baseline in the experiments.

3.4 Generated Knowledge Setup

To bypass the difficulty of matching text spans in the context to the knowledge base, we use a generative method to predict an entity e_2 in a knowledge triplet, based on the entity e_1 extracted from context $\boldsymbol{c}$ and a specified relationship r. Compared with the retrieval method described in the previous section, this method has the benefit of being able to specify the type of relation in the knowledge triplet. We can thus locate the knowledge relevant to specific tokens rather than the whole sentence. To complete the knowledge triplet, we use COMET, a framework for automatic knowledge base construction. This is a GPT model (Radford et al., 2018) finetuned on knowledge triplets from commonsense knowledge bases such as ConceptNet (Speer et al., 2017) and ATOMIC(Sap et al., 2019). The model takes $\epsilon_j = (e_1, r, *)$ as input and predicts e_2 to complete the knowledge triplet. We use the original implementation[6] and the pretrained weights on ConceptNet.

For each utterance x_i in the dialogue context, we use constituency parsing (Kitaev and Klein, 2018) to find the verb phrase and the noun phrase at depth one in the dependency tree, and use them as the input to the COMET model. Following the categorization in (Hwang et al., 2021), we limit the relationships to the commonsense subset to reduce noise and to limit the number of generated knowledge triplets. For noun phrases, the relations are mostly about their physical properties, such as *UsedFor* and *CapableOf*. For verb phrases, we focus on the social-interaction or event-centered aspects, which include relations such as *Causes* and *MotivatedByGoal*. For example, for the triplet $(loseweight, HasPrerequisite, *)$ the model predicts e_2 to be *Eat less* or *Eat healthier*.

A potential drawback of appending the knowledge at the beginning of the input is that we are not able to include information about knowledge locality as we can not tell the model which piece of the context the knowledge is corresponding to. Therefore, we take inspiration from K-BERT (Liu et al., 2020) and adopt their representation method into our BART-based model, which is referred as K-BART. We experiment with two ways to keep the structure information.We use BART-large as

[6]https://github.com/atcbosselut/comet-commonsense

the baseline, and test inserting r and e_2 without modifying the attention and positional embedding noted as *inplace*.

Soft Positional Encoding. As BART's transformer layers follow the implementation of RoBERTa, it uses a learned positional embedding, which assigns a unique embedding vector to each location in the input and captures the sequential nature of the input. For COMET generated knowledge, we plug in r and e_2 next to its corresponding e_1 in the original context. Note that the input sentence is no longer a natural sentence, which is different from instances in pretraining. Consider the following sentence with corresponding knowledge in brackets: "I've been smoking [causes cancer] too much,". This is usually regarded as two sentences: the original input "I've been smoking too much" and the introduced knowledge "smoking causes cancer." However, plain positional encoding scheme is not enough to represent this information. Hence, we treat the input sequence as a tree structure, where the r and e_2 are treated as a branch to the original input at the location next to e_1. In this case, "causes" and "too" are both considered as the fourth token right after "smoking." With this approach, the main body of the sentence will have the same index as a sentence without additional knowledge.

Mask-Self-Attention. The information introduced by a COMET generated knowledge triplet is only relevant to the first argument e_1 from the original context. Therefore, we use attention mask to modify the visibility of each part in the input sequence, and hide the introduced knowledge from other irrelevant parts of the input. The tokens in the dialogue context can see each other as usual, but the introduced knowledge r and e_2 are only visible to their corresponding e_1, which means their attention weights are always 0 for other parts of the input. In this way, unrelated tokens will not be affected by the semantics of introduced knowledge.

4 Experiments

We choose BART as the backbone network for our generation model. It is a standard seq2seq style transformer which achieved SoTA on multiple down stream tasks with a bidirectional encoder and a left-to-right decoder, which generalizes both GPT2 and BERT. Each model is trained with three random seeds.

4.1 Dataset

We use the dataset from (Pérez-Rosas et al., 2016) on Motivational Interviewing for language model fine-tuning. The dataset consists of 277 counseling sessions, covering different topics on behavior change, including smoking cessation and weight management. It has annotations on counselor verbal behaviors, such as asking a question, making a reflective response, or seeking collaboration. In the experiments, we form data samples with a reflective response as the target text $\boldsymbol{y}$ and use five former utterances within the counseling dialog as the context $\boldsymbol{c}$. That leaves us over 3000 samples after filtering.

We use ConceptNet as the knowledge base providing commonsense knowledge. It has over 21 million knowledge triplets with a set of 34 relations covering a wide variety of knowledge, including attributional relationships, causal relationships, etc. We only keep triplets that are in English and from a selected subset of relationships based on their semantic meanings, refer to the appendix for details. This leaves us with a collection of about 3.4 million triplets.

4.2 Evaluation

We evaluate our model with several common metrics. We measure the word-overlapping based relevance using BLEU-1/2 (Papineni et al., 2002), ROUGE-1/2 (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). We measure the contextual embedding similarity using BertScore (Zhang et al., 2019). We measure the diversity with the ratio of unique unigrams or bigrams among generated sentences (Li et al., 2016).

4.3 Results of Retrieval Methods

We first examine how the knowledge from different retrieval methods benefits the system. All the experiments use domain-specific knowledge as the data source. Table 2 shows our experimental results.

The *retrieval-each* method using sentence-level embeddings exceeds the baseline on Rouge-1 and METEOR, while the *retrieval-average* method, using context-level embeddings of less granularity, outperforms other methods in BLEU-2, Rouge-2, and BertScore. Meanwhile, the oracle method *retrieval-diff* unsurprisingly gets the highest score in all metrics by a large margin except Dist-1. Overall, results indicate that it is feasible to find relevant information from a domain-specific knowl-

Model	BLEU-1	BLEU-2	Rouge-1	Rouge-2	METEOR	BertScore	Dist-1	Dist-2
Baseline	**11.67**	1.38	18.94	3.04	8.90	85.39	0.21	**1.90**
Retrieval-each	10.68	1.16	**20.33**	2.99	**9.13**	85.36	0.19	1.73
Retrieval-avg	11.60	**1.43**	18.69	**3.28**	8.30	**85.44**	**0.22**	1.89
Retrieval-diff	13.63	1.80	24.23	5.24	11.41	85.99	0.21	2.01

Table 2: Performance of different retrieval methods to obtain relevant knowledge. The highest scores (excluding *Retrieval-diff* which uses grounded truth responses) for each metric are in bold.

Model	BLEU-1	BLEU-2	Rouge-1	Rouge-2	METEOR	BertScore	Dist-1	Dist-2
Baseline	11.67	1.38	18.94	3.04	8.90	85.39	0.21	1.90
Inplace	**15.40**	**3.31**	19.83	5.25	9.89	86.41	0.30	**2.43**
+Att	15.28	3.30	**20.45**	**5.50**	**10.37**	**86.50**	**0.30**	2.41
+Pos	14.79	2.73	18.83	3.72	9.25	86.31	0.30	2.29
+Att&Pos	14.24	2.54	19.28	4.01	9.52	86.19	0.29	2.32

Table 3: Experiments using soft positional encoding and masked self-attention with COMET generated knowledge.

Model	BLEU-1	BLEU-2	Rouge-1	Rouge-2	METEOR	BertScore	Dist-1	Dist-2
Baseline	11.67	1.38	18.94	3.04	8.90	85.39	0.21	*1.90*
Domain-specific	13.63	1.80	24.23	5.24	11.41	85.99	0.21	**2.01**
Commonsense	15.63	2.81	**29.19**	**7.97**	14.28	86.46	0.21	1.79
Domain & Commonsense	**16.68**	**3.03**	28.67	7.90	**14.78**	**86.66**	0.20	1.87

Table 4: The effect of using different knowledge resources during the retrieval stage.

Model	BLEU-1	BLEU-2	Rouge-1	Rouge-2	METEOR	BertScore	Dist-1	Dist-2
Commonsense	15.63	2.81	29.19	7.97	14.28	86.46	0.21	2.01
-Attribution	16.19	2.88	28.90	7.76	14.45	86.50	0.21	1.92
-Causal	**15.02**	2.56	28.60	7.31	13.91	**86.43**	0.19	1.80
-Comparison	15.66	2.71	28.34	7.23	13.96	86.47	0.23	1.99
-Conditional	16.21	2.94	28.50	7.58	14.06	86.52	0.21	1.93
-Dbpedia	15.46	2.63	27.89	7.23	13.60	86.45	0.19	**1.67**
-Intentional	15.50	**2.50**	**26.69**	**6.17**	**13.00**	86.47	0.20	1.84
-Spatial	15.48	2.64	28.79	6.92	13.99	86.46	**0.18**	1.67
-Temporal	15.80	2.63	29.14	7.33	14.56	86.43	0.19	1.71

Table 5: Ablation study for commonsense relationships. *-Causal* means a version of ConceptNet with *Causal* relationships removed is used during the retrieval. All the experiments are using the *retrieval-diff* method. The **lowest** score for each of the metrics is in bold.

edge base to improve generation given the ground truth.

4.4 Result on K-BART Model Architecture

Next, we investigate whether knowledge from COMET, a generative approach, can provide additional context to the generation task. We also evaluate whether masked attention *Att* or soft positional encoding *Pos* are better strategies to infuse knowledge by providing locality information of what tokens the knowledge is related to. We show the results in Table 3.

The *inplace* method, which inserts the relation r and the generated e_2 next to e_1, shows a significant improvement over the baseline. More specifically, the improvement in Dist-1/2 suggests that commonsense stored in COMET can also be leveraged to introduce new words and concepts into the response. Using masked attention provides further improvements in several automatic metrics, except for a slightly lower BLEU score. Interestingly, the soft positional encoding worsens the performance regardless being used by itself or when combined with masked attention. One potential explanation for this is that BART is more robust to masked attention as its effects are similar to attention dropout, while the soft positional encoding causes more position collision and requires more training samples to be effective.

4.5 Experiments Varying Knowledge Source

After showing that both retrieved and generated knowledge helps to improve the generation of counseling responses, a natural question that follows is: how does the knowledge resource itself affect the overall performance?

To explore this question, we conducted a set of comparative experiments on using domain-specific

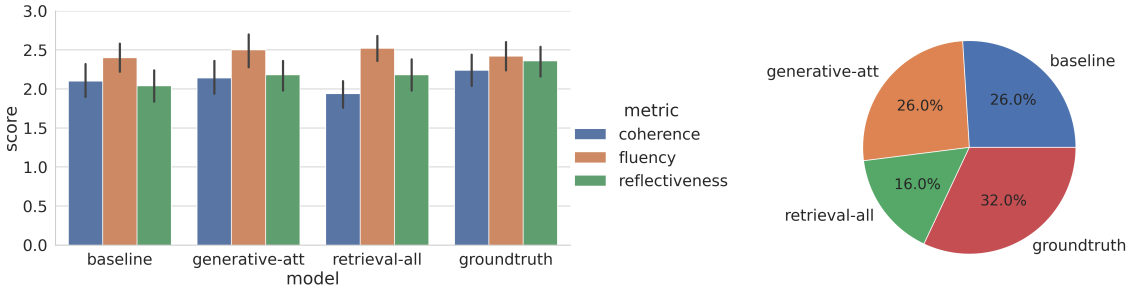

Figure 4: Left: the average scores on human evaluation metrics. Right: the percentage being chosen as the best responses generated by different models. *generative-att* model is the generative setup with only masked attention applied. *retrieval-all* is the retrieval setup where both domain knowledge and commonsense are available.

and commonsense knowledge. During our experiments, we use the *retrieval-diff* method, which can be seen as an upper-bound of performance using the actual ground truth response. The knowledge candidates are obtained from either ConceptNet triplets in their surface text form or domain-related knowledge collected from the Internet as described in §3.2.

Domain Specific Knowledge vs Commonsense Knowledge. As shown in Table 4 both domain-specific knowledge and commonsense knowledge serve as useful sources of knowledge resources for our generation task. However, the model using ConceptNet performs significantly better than the model using domain-specific knowledge in all metrics except Dist-2. One potential reason for this is that the sheer amount of commonsense knowledge is much larger than the amount we collect and has better coverage for what is mentioned in the dialogue context. However, our experiments show that aggregating both types of knowledge further improves the system's performance. This suggests the domain-specific knowledge provides complementary information relevant to the counseling domain, such as the side effect for a medication, that is not captured by the commonsense knowledge base. Note that more than 20% of the retrieved sentences are from the domain-specific knowledge base, while the commonsense knowledge base is more than 30 times larger in size. This further shows that our data collection pipeline is able to provide knowledge that is more relevant to the dialogue context, with the added benefit of no human annotation involved.

The Role of Different Types of Commonsense Knowledge. We evaluate the role of different types of knowledge by conducting an ablation study based on the main categories in Conceptnet, including attribution, causal, comparison, conditional, intentional, spatial, and temporal categories.

We build separate models by removing a commonsense knowledge category at a time. Results in Table 5 show that removing the *intentional* relationships harms the performance the most on Rouge-1/2 and METEOR, and removing the *causal* relationships leads to the lowest score on BLEU-1 and BertScore. Interestingly, these relations are important for counseling conversations where the counselor usually infer the intention or causes behind their clients statements. For instance, in smoke cessation counseling, counselors might be aware that the main reasons to quit are related to well-being or personal relationships.

Removing a few sets of relationships, such as *Attribution* or *Temporal*, causes minimal performance drop or even an improvement. These results suggest that those relationships are not salient or introduce noise during the retrieval process.

4.6 Human Evaluation

We conduct a human evaluation where we ask annotators to indicate their preferences between our best performing models from both the retrieval and the generative settings, and a model without knowledge enhancement. We evaluated each each model response using three metrics: **Fluency** indicating whether the sentence is grammatically correct and natural; **Coherence** indicating whether the response is on topic and relevant to the dialogue history; **Reflectiveness** indicating if the response summarizes what the client has said or interprets what the client means. All these metrics are scored with a three-point Likert scale.

We also ask the annotators if the retrieved knowledge is helpful for generating a better response, where the knowledge is triplets for the generative setup and sentences for the retrieval setup. In addi-

tion, we ask the annotators to pick the best response between our models and the ground truth.

We randomly choose 50 samples for each model to be annotated. The annotation was conducted by two annotators using Qualtrics.[7] The annotators had no information on which model generated the the response being annotated.

Figure 4 shows the average score for each metric and the percentage of times each system was chosen as the best response. Results show that the ground truth responses have the highest score in terms of *reflectiveness* and *coherence*. A potential reason for this is that the ground truth responses are generally longer, thus containing more information from the dialogue context. As for the best response, the ground truth was also the most picked one and our models using knowledge have not outperformed the baseline in this regard.

The model using generated knowledge triplets outperforms the baseline in all three metrics, suggesting the *motivation* and *cause* relationships generated by COMET brought useful context to the dialog. However, only 22% of the triplets sampled from the test set are considered helpful by our annotators. This calls for closer inspection on the difference between how the models take advantage of commonsense knowledge and how humans perceive it. The model using retrieved knowledge assertions outperforms the baseline on fluency and reflectiveness but has a low coherence score. Among the knowledge assertions, 38% of retrieved sentences are relevant to the dialog when using domain knowledge, and 48% for commonsense knowledge.

5 Conclusion

In this paper, we proposed the task of knowledge enhanced counseling reflection generation, and experimented with different ways to introduce knowledge into the reflection generation model using both retrieval and generative settings. We found that both strategies benefit the generation task on various automatic metrics, which is further consolidated by the human evaluation. In addition, we showed that counseling domain knowledge serves as good complementary knowledge source to ConceptNet. Through an ablation study, we found that commonsense related to *intentional* and *causal* relationships is essential for the counseling domain.

[7] www.qualtrics.com

Acknowledgments

We are grateful to Kenneth Resnicow and Larry An for their expert input on the importance of domain knowledge in counseling. This material is based in part upon work supported by the Precision Health initiative at the University of Michigan, by the National Science Foundation (grant #1815291), and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Precision Health initiative, the National Science Foundation, or John Templeton Foundation.

References

Fahad Almusharraf, Jonathan Rose, and Peter L. Selby. 2020. Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing–based chatbot through iterative interactions. *Journal of Medical Internet Research*, 22.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Nam Do and Ellie Pavlick. 2021. Are rotten apples edible? challenging commonsense inference ability with exceptions. In *FINDINGS*.

Kathleen Kara Fitzpatrick, Alison M Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depres-

sion and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, 4.

Stephanie Greer, Danielle E. Ramo, Yin-Juei Chang, Michael Fu, Judith Tedlie Moskowitz, and Jana Haritatos. 2019. Use of the chatbot "vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. *JMIR mHealth and uHealth*, 7.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid 'O S'eaghdha, Sebastian Pad'o, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Elizabeth M Huffman, Dimitrios I Athanasiadis, Nicholas E Anton, Lindsay A Haskett, Dominique L Doster, Dimitrios Stefanidis, and Nicole K Lee. 2021. How resilient is your team? exploring healthcare providers' well-being during the covid-19 pandemic. *The American Journal of Surgery*, 221(2):277–284.

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*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *ACL*.

Neil Katz and Kevin McNulty. 1994. Reflective listening.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. 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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*.

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. *arXiv preprint arXiv:1907.11692*.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. Improving neural story generation by targeted common sense grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5988–5993, Hong Kong, China. Association for Computational Linguistics.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mario R Paredes, Vanessa Apaolaza, Cristóbal Fernandez-Robin, Patrick Hartmann, and Diego Yañez-Martinez. 2021. The impact of the covid-19 pandemic on subjective mental well-being: The interplay of perceived threat, future anxiety and resilience. *Personality and Individual Differences*, 170:110455.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019a. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019b. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8697–8704.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Siqi Shen, Charles F Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *SIGDIAL*.

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, Online. Association for Computational Linguistics.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *IJCAI*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Belfin R V, Shobana A J, Megha Manilal, Ashly Mathew, and Blessy Babu. 2019. A graph based chatbot for cancer patients. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 717–721.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) 2020*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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.

Category	Relations	Surface Text	Example
Attribution	Capable Of	is capable of	knife → cut
	Part Of	is a part of	gearshift → car
	Not Capable Of	is not capable of	
	Has A	has a	bird → wing; pen → ink; gearshift → car
	Has Property	is	ice → cold
	Not Has Property	is not	
	Has Subevent	includes	eating → chewing
	Is A	is a	car → vehicle; Chicago → city
	Manner Of	is a manner of	auction → sale
	Instance Of	is an instance of	
	Made of	is made of	bottle → plastic
	Symbol Of	is a symbol of	red → fervor
Causal	Causes	causes	exercise → sweat
	Causes Desire	makes someone want to	having no food → buy food
Comparison	Antonym	is an antonym of	black →white; hot → cold
	Distinct From	is distinct from	red → blue; August → September
	Similar To	is similar to	mixer → food processor
	Synonym	is a synonym of	sunlight → sunshine
	DefinedAs	is defined as	peace → absence of war
	DerivedFrom	is derived from	pocketbook → book
	FormOf	is a form of	slept → sleep
Conditional	Has Prerequisite	requires	dream → sleep
	Entails	entails	
	Has Context	has a context of	astern → ship; arvo → Australia
Intentional	Desires	desires	person → love
	Not Desires	does not want	
	Motivated By Goal	is motivated by	compete → win
	Used For	is used for	bridge → cross water
	Receives Action	can be	button → push
Spatial	At Location	is usually located at	try clothes → changing room
	Located Near	is usually located near	table → chairs
Temporal	Has First Subevent	begins with	sleep → close eyes
	HasLastSubevent	concludes with	cook → clean up kitchen

Table 6: Relations in the ConceptNet grouped into categories

Category	Count
Attribution	320495
Causal	21487
Comparison	981216
Conditional	256047
Intentional	61367
Spatial	27844
Temporal	6219
Dbpedia	12295
Total	3422957

Table 7: Number of knowledge triplets in each relation category.