Improving Mental Health Support Response Generation with Event-based Knowledge Graph

Lingbo Tong,[1] Qi Liu, [2] Wenhao Yu [1] Mengxia Yu [1] Zhihan Zhang [1] Meng Jiang [1]

[1] University of Notre Dame, [2] University of Washington
{ltong2, wyu1, myu2, zzhang23, mjiang2}@nd.edu, qliu3@uw.edu

Abstract

Online mental health support plays a significant role in people's well-being and suicidal intervention. In this paper, we propose the task of mental health support response generation for online forums, which is a challenging task to existing work focusing on empathetic conversational systems. To bring in external knowledge for the proposed task, we construct a knowledge graph, MHKG [1], which consists of eventualities from ASER and contextual relations from Reddit corpus. We conduct experiments that leverage MHKG for text generation. Both automatic and human evaluation results suggest that enriching the input sequence with the ground-truth neighbors in MHKG is able to significantly improve model performance. We thus propose inference on MHKG to find ground-truth neighbors as the future direction.

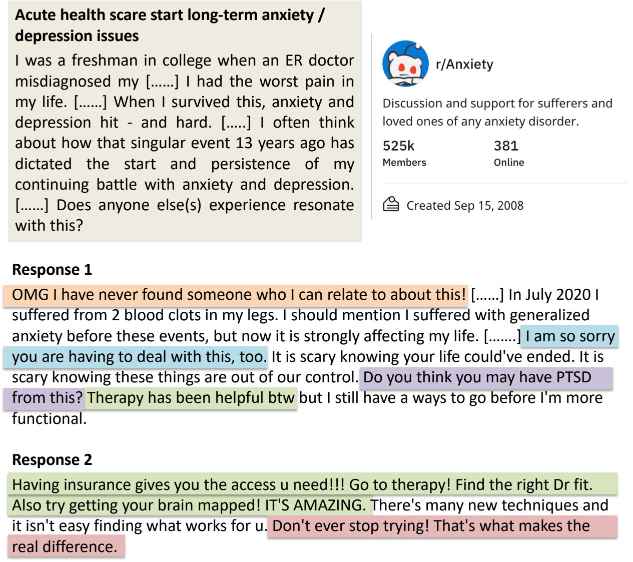
Introduction

Today's social media platforms have continuously emerged as important vehicles for the general population to prevent or save themselves from mental illness. Online mental health support plays a significant role in helping people who suffer from school or workplace bullying, toxic interpersonal relationships, etc. Studies suggest that supportive interactions through social media are associated with lower depression-related thoughts and can be protective against suicidal behaviors (Choi and Noh 2020; Cole et al. 2017).

While supportive responses can be quite beneficial, many help-seeking posts in online forums, such as Reddit, often cannot get in-time responses; or even worse, they get no response at all. So, an intelligent system with the knowledge to generate supportive responses can be significant for mental health support, especially for self-harm intervention.

In this paper, we address the task of generating mental health support responses for online forums. Specifically, given a user post on Reddit, an intelligent system is expected to generate a response that provides mental health support.

Figure 1 shows a post from the Reddit anxiety channel. Generating mental health support responses for online forums posts two unique challenges. First, unlike dialogues that are usually multi-round and short-formed, most posts and responses on social media forums are long and richer in content (Ma et al. 2020). Second, these responses often contain diverse strategies and rich external knowledge. Therefore, recent works investigating empathetic or emotional support conversation systems (Cheng et al. 2022; Peng et al. 2022) face severe challenges when applied to our proposed task, as they are lack of external knowledge for generating long and informative responses.



Acute health scare start long-term anxiety / depression issues

I was a freshman in college when an ER doctor misdiagnosed my [......] I had the worst pain in my life. [......] When I survived this, anxiety and depression hit - and hard. [.....] I often think about how that singular event 13 years ago has dictated the start and persistence of my continuing battle with anxiety and depression. [......] Does anyone else(s) experience resonate with this?

r/Anxiety

Discussion and support for sufferers and loved ones of any anxiety disorder.

525k Members

381 Online

Created Sep 15, 2008

Response 1

OMG I have never found someone who I can relate to about this! [......] In July 2020 I suffered from 2 blood clots in my legs. I should mention I suffered with generalized anxiety before these events, but now it is strongly affecting my life. [.......] I am so sorry you are having to deal with this, too. It is scary knowing your life could've ended. It is scary knowing these things are out of our control. Do you think you may have PTSD from this? Therapy has been helpful btw but I still have a ways to go before I'm more functional.

Response 2

Having insurance gives you the access u need!!! Go to therapy! Find the right Dr fit. Also try getting your brain mapped! IT'S AMAZING. There's many new techniques and it isn't easy finding what works for u. Don't ever stop trying! That's what makes the real difference.

Figure 1: A post from Reddit anxiety (r/anxiety) channel and two responses. Sensitive information has been removed.

Plenty of works have been proposed to incorporate large-scale commonsense knowledge graphs as external knowledge sources, including ConceptNet (Speer, Chin, and Havasi 2017), ATOMIC (Sap et al. 2019), and ASER (Zhang et al. 2020a). While these knowledge graphs are useful in generating responses for open-domain questions, their psychology-related sub-graphs are extremely sparse, indicating their limited expertise in the field of psychology, particularly in mental health support.

To tackle these challenges, we propose **M**ental **H**ealth **K**nowledge **G**raph (MHKG), in which its nodes are eventu-

[1] https://github.com/Stan7s/MHKG

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.

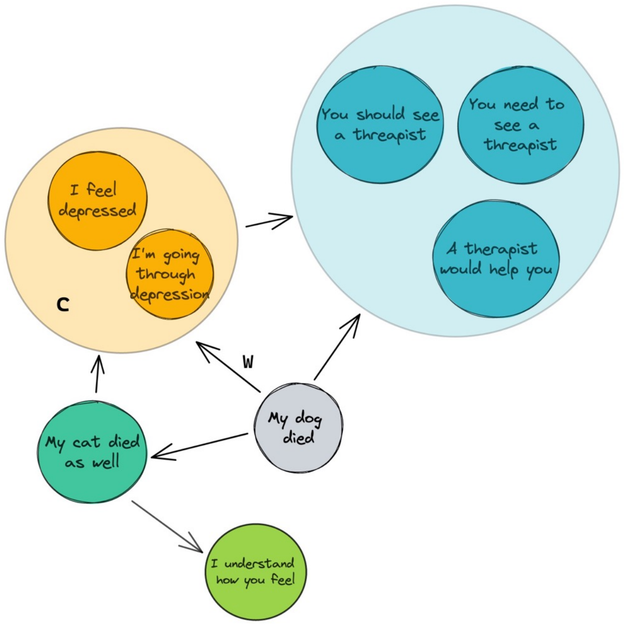

You should see a threapist
You need to see a threapist
I feel depressed
I'm going through depression
A therapist would help you
C
W
My dog died
My cat died as well
I understand how you feel

Figure 2: An example of MHKG.

alities extracted from ASER (Zhang et al. 2020a), and edges are relations between eventualities, as shown in Figure 2. We incorporate our MHKG in the downstream task of generating mental health support responses. We take it as an external knowledge source and use its sub-graphs to enrich the input text (i.e., an user post). Our evaluation shows that sub-graphs containing ground-truth neighbors are useful in generating better response, under both automatic and human evaluation, especially in *relevance* and *coherence*.

Our main contributions can be summarized as follows: 1) we formulate the task of online mental health response generation; 2) we propose to construct MHKG, a novel mental health support knowledge graph that incorporates existing large eventuality knowledge graph and contextual knowledge from textual corpus; 3) we evaluate the effectiveness of integrating the proposed MHKG in generating mental health support responses, and discuss possible approaches of utilizing it for text generation as our next step.

Related Work

Empathatic Conversation System

A lot of existing works have attempted to build empathetic conversation systems (Raamkumar and Yang 2022). For instance, Sabour, Zheng, and Huang (2022) propose a Commonsense-aware empathetic response generation system. Tu et al. (2022) utilize user's fine-grained emotional status and respond with a mixture of strategies. Li et al. (2022) leverage external knowledge to understand and express emotions through constructing a emotional context graph. Unlike these works focusing on predicting emotion-related labels, we propose to leverage contextual knowledge that contains richer information through MHKG.

Table 1: Statistics of the Reddit corpus collected.

SubReddit	**# posts**	**# sents.**	**# sents. in posts**	**# sents. in replies**
lifesucks	85	225	1,321	1,304
emotionalsupport	490	1,016	7,252	6,662
psychotherapy	748	7,745	8,076	43,778
sad	6,080	16,771	73,266	91,852
psychonaut	7,369	77,146	123,160	420,236
anxiety	47,547	158,533	596,752	913,396
suicidewatch	104,652	308,454	1,438,198	1,719,802
depression	163,787	523,771	2,191,421	2,909,972
offmychest	175,646	661,554	3,324,618	3,446,981
all	**506,404**	**1,755,215**	**7,764,064**	**9,553,983**

KG-Enhanced Text Generation

External knowledge is vital for understanding and generating informative responses. Recent work have proved that, with the help of encoding knowledge into input texts, the generated texts could potentially incorporate more meaningful contents along with other commonsense references to input sentences (Yu et al. 2022). Different ways of utilizing knowledge graphs for open-domain generation tasks have been investigated. For example, several recent studies proposed methods for inferring multihop paths on relational knowledge graph through graph neural networks (GNN) (Ji et al. 2020; Yu et al. 2021; Ju et al. 2022).Ji et al. (2020) used multi-relational paths from the external commonsense KG for dynamic multi-hop reasoning. Ju et al. (2022) incorporated neighbors in Wikidata in the context representations of the input sequence by employing graph attention.

Dataset Collection

We collected the dataset from Reddit using the Reddit API. A total of 500k posts and 1.7M responses were collected from 10 SubReddit channels that contains mental health related self-disclosures and supportive responses from peer users. Table 1 shows the statistics of each SubReddit channel. We then adopt NLTK Tookit for data cleaning, including tokenizing and removing informal tokens like abbreviations, emoji and web urls.

Proposed Method

The goal of our present work is to develop a KG-augmented pipeline for online mental health response generation. In this section, we first formulate the task of mental health support response generation. Then, we introduce our text corpus collected from Reddit mental health channels. Next, we introduce the construction of MHKG that incorporates knowledge from both the corpus and ASER.

Problem Formulation

Suppose a help-seeking post P consists of m sentences, denoted as $P = [p_1, \cdots, p_m]$. Similarly, the response R that addresses the concern of the post and provide mental health support is denoted as $R = [r_1, \cdots, r_n]$. Our proposed task is then defined as follows: Given a original post P and the previous k sentences of it corresponding response R, i.e., $[r_{max\{i-k+1,1\}}, \cdots, r_i]$ where $1 <=$

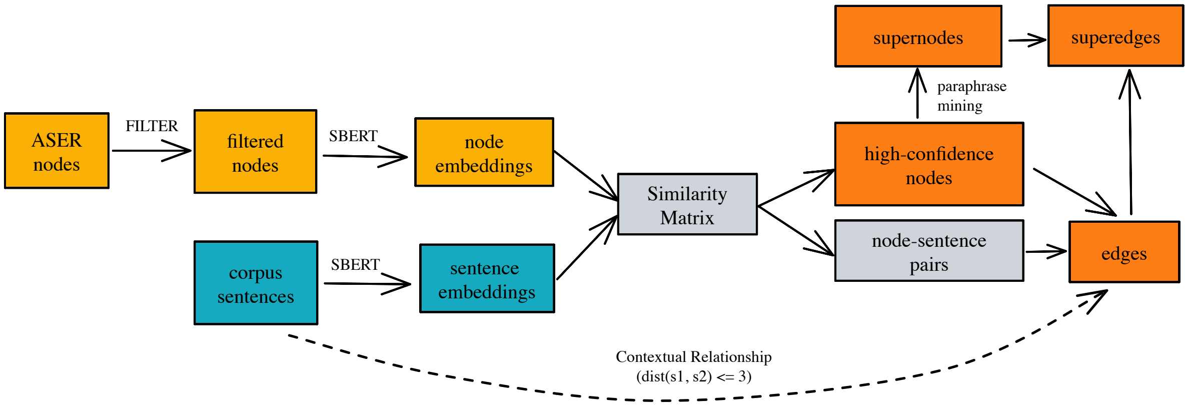


Figure 3: Build KG.

$i < n$, generate the $(i+1)$-th sentence of the target response r_{i+1}, i.e., learning a language generation model $f_\theta : (P, r_{max\{i-k+1,1\}}, \cdots, r_i) \rightarrow r_{i+1}$, parameterized by θ.

MHKG Construction

We denote our proposed knowledge graph as $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where the $\mathcal{V} = \{V_i\}$ is a set of n supernodes ($i = 1, \cdots, n$), and $\mathcal{E} = \{E_i\}$ is a set of m superedges ($i = 1, \cdots, m$). Each supernode $V_i = \langle \{X_{k_i}\}, c_i \rangle$, where X_i is a set of k synonymous eventualities and c_i is a confidence score. Each superedge $\{E_i\} = \langle X_{i1}, w_i, X_{i2} \rangle$, where X_{i1} and X_{i2} are two eventualities and w_i is the weight of the superedge. Figure X shows the overall structure of MHKG.

The process of building MHKG is shown in Figure 3. First, we collected eventualities from ASER. Then we filter these eventualities with psychological-related keywords followed by a further selection with similarity-based confidence scores. After that, we build edges among eventualities based on contextual relationships extracted from the corpus. Finally, we merge synonymous nodes and get the final knowledge graph. The detail of each step is introduced in the following subsections.

Eventuality Nodes

The eventualities in our knowledge were gathered from ASER (Zhang et al. 2022), a large-scale weighted eventuality knowledge graph. Our motivation of adopting ASER is two fold. On one hand, compared with direct extraction from corpus, adopting the huge amount of eventualities from ASER can greatly help to enrich our knowledge graph on the semantic level with a low computational cost. On the other hand, ASER contains limited contextual knowledge under the mental health domain. As is shown in Table 2, the subgraph of ASER contains 178K mental health-related eventualities yet only contains 8K edges. Therefore, we choose to make use of only the eventualities from ASER and build edges among them based on the contextual information from our self-collected Reddit mental health corpus.

Filter eventualities with psychological keywords. The goal of filtering was to collect eventualities that are meaningful in the context of mental health support. To achieve that, we first constructed a keyword list. The list contains 1,304 keywords, composed of 1) words from mental health related categories in LIWC dictionary and 2) words that appear more than 1,000 times in our Reddit corpus. Stopwords and common words were excluded from the list. After that, we filtered all eventualities in ASER with the keyword list and only kept eventualities containing at least two keywords. We also filtered out eventualities that contains less than 3 words or more than 9 words to assure the completeness and representativeness of the candidates.

Confidence-based Ranking. After obtaining the eventuality candidates, we further ranked them based on embedding similarities. First, we adopted SentenceBERT (Reimers and Gurevych 2019) to encode both eventualities and corpus sentences. Specifically, we chose the model 'ALL-MINILM-L6-V2'. Then, we computed cosine similarity scores $\text{sim}(v_i, s_j)$ between embedding of eventuality v_i and embedding of sentence s_j in the corpus. For each eventuality, we identified its top-k (we set $k = 10$ in the experiments) similar sentences based on the similarity matrix, and took the median of the top-k similarity scores as its *confidence score*. Higher confidence score indicates that the eventuality is more likely to involve mental health topics. For example, "you need to see a therapist" has a confidence score $\text{sim}(v_i, s_j) = 0.97$, while "the car was manufactured 10 years ago" only results in a score $\text{sim}(v_i, s_j) = 0.43$. We ranked these eventualities by their confidence scores and only kept those with scores $\text{sim}(v_i, s_j) \geq 0.6$, resulting in a total of 178k high-confidence eventualities. These eventualities are regarded as nodes for our knowledge graph.

Contextual Relations

In the previous step, we paired eventuality nodes and sentences in the corpus if similarity scores $\text{sim}(v_i, s_j) \geq 0.6$. Now, we create an edge between the eventuality node pairs (v_i, v_j) if sentences in the corpus paired with v_i and v_j are adjacent. Specifically, we consider two sentences to be adjacent if they are in the same reply document with less than three sentences in between. The weight of an edge is the

Table 2: Descriptive statistics of MHKG.

	# Nodes (k)	# Edges (k)	Avg. Node Degree	# Supernodes (k)	# Superedges (k)	Avg. Supernode Degree
ASER(core)	53,000	52,000	0.98	-	-	-
ASER (core) ∩ Corpus	178	8	0.04	-	-	-
MHKG (ours)	178	60,658	340.78	121	34,626	286.17

number of times it was added during the calculation. The intuition behind this approach was that sentences that appeared close in text might share latent contextual relations. By linking their corresponding eventualities together, we stored the contextual information in an explicit way, which could be considered as a kind of knowledge augmentation.

Synonymous Nodes

After the steps above, a basic version of our KG has been generated. However, both ASER and the corpus contain a large amount of paraphrases, i.e., events or sentences that are not exactly the same on the token level but with duplicated semantic meanings. To solve this issue, we adopted a paraphrase mining approach to merge synonymous nodes together. Specifically, for a pair of nodes (v_i, v_j), we merged them together if the following criteria were satisfied:

$$\mathrm{sim}(\mathrm{SBERT}(v_i), \mathrm{SBERT}(v_j)) \geq 0.9$$

$$\mathrm{sim}(\text{TF-IDF}(v_i), \text{TF-IDF}(v_j)) \geq 0.6$$

Conducting this process on every pair of nodes in the graph resulted in communities composed of synonymous nodes, which are called *supernodes*. Similarly, the edge between two supernodes is called *superedge*, which was aggregated based on the original edges between nodes in both supernodes.

Experiments

Experimental Setup

We use DialoGPT (Zhang et al. 2020b) as our backbone model. DialoGPT is a tunable gigaword scale GPT-2 model for generation of conversational reponses, trained on Reddit data. It excels at ranking potential responses and handling informal textual data on social media. Specifically, we adopt the pre-trained model DialoGPT-small from Huggingface.

We leverage MHKG by incorporating its subgraph into the input sequence. Specifically, we evaluate five models with different input. Suppose we aim to generate the i-th sentence of a response. Model 1 (M1) only takes the previous k sentences in this response as input. Model 2 (M2) takes both 1) the previous k sentences and 2) the top 10 neighbors (i.e., 1-hop supernodes) of the eventuality in MHKG that has the highest similarity with the $(i-1)$-th sentence. The neighbors are ranked by the product of edge weight and the sum of confidence scores of both nodes. Model 3 (M3) is similar to M2, only that it changes 10 neighbors to 20. Model 4 (M4) concatenates 1) the previous k sentences, 2) the top 9 neighbors, and 3) the ground-truth neighbor, i.e., the neighbor that has a high similarity with the $(i-1)$-th sentence. Model 5 (M5) concatenates only the previous k sentences and the ground-truth neighbor. Figure 4 illustrates the relationships among sentences and nodes in MHKG.

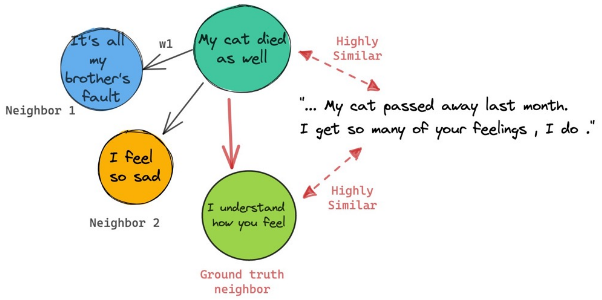

It's all my brother's fault
w1
My cat died as well
Highly Similar
Neighbor 1
"... My cat passed away last month.
I get so many of your feelings , I do ."
I feel so sad
Highly Similar
Neighbor 2
I understand how you feel
Ground truth neighbor

Figure 4: Illustration of relationships among sentences and nodes in MHKG when building training dataset.

In terms of training data, we take a subset of Reddit corpus containing a total of 40,000 samples, and split it into train/val/test as 8:1:1. The three sets do not overlap with each other. Furthermore, our validation and test set are independent from the building process of MHKG to avoid data leakage.

Evaluation Metrics

Automatic Metrics. They are used for evaluating the correspondence between the predicted output and the model output. We adopt several wildly-used evaluation metrics following recent work in empathetic generation, including BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). Specifically, we report BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L F1, and CIDEr.

Human Evaluation. We recruit four human annotators, three of whom are experienced in NLP and one in Psychology, and ask them to rate the generated responses according to five aspects. These aspects include three metrics including *fluency*, *coherence*, and *empathy*, introduced by Rashkin et al. (2018), and two original metrics, *coherence* and *supportiveness*. Table 3 shows the metrics and corresponding questions. Each aspect is rated with levels from 1 to 5. The expert annotators do not know which model the response is before all annotations were finished.

Analysis

Automatic Evaluation Results

The automatic evaluation results are shown in Table 4, from which we can draw the following conclusions:

Ground-truth neighbors are effective in guiding generation. M5 achieved leading performance among the five models, as expected, followed by M4. This proves that prompts

Table 3: Human evaluation metrics and their corresponding questions.

Metric	Question
Fluency	Could you understand the generated sentence? Did the language seem accurate?
Revelence	Did the generated sentence seem appropriate to the post? Was it on-topic?
Coherence	Was the generated sentence consistent with the previous sentences in the response? Was it on-topic?
Empathy	Did the generated sentence show understanding of the feelings of the person talking about their experience?
Supportiveness	Did the generated sentence provide at least one of the following types of support: 1) encouragement, 2) problem analysis, 3) advice?

Table 4: Automatic Evaluation Results.

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
M1: r3	0.1719	0.0467	0.0202	0.0125	0.0649	0.1788	0.1273
M2: r3+top10	0.1668	0.0469	0.0227	0.0126	0.0658	0.1832	0.1288
M3: r3+top20	0.1753	0.0453	0.0164	0.0070	0.0666	0.1822	0.1209
M4: r3+top9+true	0.2930	0.1530	0.0965	0.0662	0.1412	0.3161	0.8824
M5: r3+true	**0.3478**	**0.2047**	**0.1354**	**0.0925**	**0.1824**	**0.3879**	**1.3088**

Table 5: Human Evaluation Results.

	Flu.	Rev.	Coh.	Emp.	Sup.
M1: m3	4.48	3.08	2.98	3.23	3.06
M2: m3+top10	4.55	3.23	3.08	3.11	3.10
M3: m3+top20	4.48	3.13	3.32	2.99	3.14
M4: m3+top9+true	**4.66**	**3.63**	**3.64**	**3.45**	**3.38**
M5: m3+true	4.46	3.56	**3.64**	3.41	**3.38**
Human	**4.87**	**3.79**	**4.03**	**3.63**	**3.56**

including ground-truth neighbors are effective in guiding response generation. Moreover, the outstanding performance of M4 compared to M1-3 indicates that, even if the ground-truth neighbor was mixed with other neighbors, it can still significantly improve the quality of the generated output.

Top-k neighbors are not helpful. We observe that M2 is not significantly better than M1, which indicates top neighbors are of little help in guiding generation toward the direction of the ground truth sentence. In fact, on the basis of M2, adding more neighbors (M3) leads to a decline in all automatic metric scores except BLEU_1. One possible explanation is that ground-truth neighbors are often not included in the top k neighbors, and the two are far in the semantic space. Adding more neighbors will thus bring noises rather than ground-truth knowledge as k becomes large. Further investigation needs to be done to validate this assumption.

Human Evaluation Results

Table 5 shows the human evaluation results. We observe that M4's performance is next to the original response written by human than the others in all five human-evaluated aspects, followed by M5. As a comparison, M1, the model that only uses previous text as inputs, scored the lowest in all 5 aspects. We thus infer that the ground-truth neighbor as well as a moderate number of neighbors can be useful in guiding generation.

Case Study

Table 6 shows an example of sentences generated by models and human in the test set. Compared with other results, the output of M1 is relatively short and general. The sentence generated by M2 is relevant but not coherent; plus, it missed the keyword ”anxiety”. M5, the best model in auto evaluation, produced a sentence that lacks fluency (”lost health and anxiety”). On the contrary, M3 was able to produce a long sentence that is both relevant and rich in details. Note that although M3 performed unsatisfactory in auto evaluation, in this example it also produced a fluent, relevant sentence which contains an analogy. Two hypothesis could be drawn from here: 1) The combination of ground-truth neighbors and top neighbors might be beneficial for generating responses that optimize between accuracy and diversity. 2) Better auto evaluation metrics or models need to be investigated.

Conclusion

In this paper, we define the task of mental health support response generation for online forums, which is to generate the next sentence based on previous contents in a response. To bring in external knowledge for this task, we construct a knowledge graph, MHKG, composed of eventualities and contextual relations. We conduct experiments that leverage the knowledge graph for text generation. We conclude that enriching the input sequence with the ground-truth neighbors in MHKG is able to improve model performance on our proposed task. However, we also found that simply selecting top k neighbors is not able to cover the ground-truth neighbors. To address this issue, we plan to incorporate original posts as a part of the input as our next step. We also plan to design multi-hop reasoning approach on graph to see if it can help to reach the ground-truth neighbors.

Table 6: A case of next sentences generated by models (M1-M5) and human based on the post and previous sentences in the response.

Post	Previous Sentence(s) in Response	Next Sentence
So in the last few months I have noticed that I have this feeling of dread for no reason at all . I have always felt a looming dread but it was occasionally or for a specific reason like upcoming exams. However, recently it has increased significant and multiple times a day. For example, I would being playing Sonic and all of a sudden I would be feel it hit me. I would carry on with the game hoping it would go away.	I get a random feeling of dread sometimes too.	M1: I can't escape.
		M2: I just want to know what that dread is.
		M3: Like a kind of cloud of anxiety that comes over you though.
		M4: My anxiety has been going on for a long time and I just can't seem to get enough sleep.
		M5: Like, I 've lost a lot of health and anxiety in the past 6 months.
		Human: I suffer with health anxiety.

References

Banerjee, S.; and Lavie, A. 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*, 65–72.

Cheng, Y.; Liu, W.; Li, W.; Wang, J.; Zhao, R.; Liu, B.; Liang, X.; and Zheng, Y. 2022. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. *arXiv preprint arXiv:2210.04242*.

Choi, D.-H.; and Noh, G.-Y. 2020. The influence of social media use on attitude toward suicide through psychological well-being, social isolation, and social support. *Information, communication & society*, 23(10): 1427–1443.

Cole, D. A.; Nick, E. A.; Zelkowitz, R. L.; Roeder, K. M.; and Spinelli, T. 2017. Online social support for young people: does it recapitulate in-person social support; can it help? *Computers in human behavior*, 68: 456–464.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; Zhu, X.; and Huang, M. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. *arXiv preprint arXiv:2009.11692*.

Ju, M.; Yu, W.; Zhao, T.; Zhang, C.; and Ye, Y. 2022. Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. *arXiv preprint arXiv:2210.02933*.

Li, Q.; Li, P.; Ren, Z.; Ren, P.; and Chen, Z. 2022. Knowledge bridging for empathetic dialogue generation.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Ma, Y.; Nguyen, K. L.; Xing, F. Z.; and Cambria, E. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64: 50–70.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peng, W.; Hu, Y.; Xing, L.; Xie, Y.; Sun, Y.; and Li, Y. 2022. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. *arXiv preprint arXiv:2204.12749*.

Raamkumar, A. S.; and Yang, Y. 2022. Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities. arXiv:2206.05017.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sabour, S.; Zheng, C.; and Huang, M. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11229–11237.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.-R.; and Yan, R. 2022. MISC: A MIxed Strategy-Aware Model Integrating COMET for Emotional Support Conversation. *arXiv preprint arXiv:2203.13560*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Yu, D.; Zhu, C.; Fang, Y.; Yu, W.; Wang, S.; Xu, Y.; Ren, X.; Yang, Y.; and Zeng, M. 2021. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *arXiv preprint arXiv:2110.04330*.

Yu, W.; Zhu, C.; Li, Z.; Hu, Z.; Wang, Q.; Ji, H.; and Jiang, M. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys (CSUR)*.

Zhang, H.; Liu, X.; Pan, H.; Ke, H.; Ou, J.; Fang, T.; and Song, Y. 2022. ASER: Towards Large-Scale Commonsense Knowledge Acquisition via Higher-Order Selectional

Preference over Eventualities. *Artificial Intelligence*, 309: 103740.

Zhang, H.; Liu, X.; Pan, H.; Song, Y.; and Leung, C. W.-K. 2020a. ASER: A large-scale eventuality knowledge graph. In *Proceedings of the web conference 2020*, 201–211.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. B. 2020b. 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*, 270–278.