Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context

Siqi Shen, Charles Welch, Rada Mihalcea, Verónica Pérez-Rosas
Department of Computer Science and Engineering,
University of Michigan
{shensq, cfwelch, mihalcea, vrncapr}@umich.edu

Abstract

In this paper, we introduce a counseling dialogue system that provides real-time assistance to counseling trainees. The system generates sample counselors' reflections – i.e., responses that reflect back on what the client has said given the dialogue history. We build our model upon the recent generative pretrained transformer architecture and leverage context augmentation techniques inspired by traditional strategies used during counselor training to further enhance its performance. We show that the system incorporating these strategies outperforms the baseline models on the reflection generation task on multiple metrics. To confirm our findings, we present a human evaluation study that shows that the output of the enhanced system obtains higher ratings and is on par with human responses in terms of stylistic and grammatical correctness, as well as context-awareness.

1 Introduction

A recent survey on mental and behavioral healthcare showed that while there is an increasing need for counseling services, the available mental health workforce is barely coping with this demand.[1] An important reason behind this unmet need is that the training of counselors requires a lot of time and effort. Typically, counselor training involves refining counseling skills through practice and feedback using role-play activities, simulated patients, or real patients, thus heavily relying on human supervision and interaction.

In clinician training, feedback and coaching can significantly improve the post-training counselor proficiency (Miller et al., 2004). However, the standard way of providing systematic feedback relies on human coding of the counseling sessions. This process can take up to ten times as long as the duration of the session itself, and thus it does not scale up (Atkins et al., 2014).

Previous work has focused on developing automatic tools for counseling evaluation and training tasks, including automatic coding (i.e., recognizing a counselor behavior) and forecasting (i.e., predicting the most appropriate behavior for the next counselor's utterance) (Tanana et al., 2016; Park et al., 2019; Cao et al., 2019). These tools aim to facilitate the evaluation of a counseling encounter and, to some extent, provide generic guidance during the conversation. Although these systems help counselors by suggesting the timing of a certain counseling behavior, they do not offer any help on how to acomplish it.

Among the different skills to be learned by counselors, reflective listening has been shown to be an important skill related to positive therapeutic outcomes (Moyers et al., 2009). Reflective listening is a conversational strategy used by counselors to show that they understand their clients' perspectives, feelings, and values (Miller and Rollnick, 2013). During this process, the counselor listens to the client's statements and then makes a statement (reflection) that is a reasonable approximation of the meaning of what the client has said. Thus, the main role of reflections is to keep the conversation focused on the client and to move the conversation forward. For example, considering the following utterance by the client, a counselor could make reflections (a) or (b) to show an understanding of the client's feelings and concerns.

> *Client:* I want to quit smoking because I don't want another heart attack; I want to see my kids grow up.
> *Counselor (a):* You are scared that you might have another heart attack.
> *Counselor (b):* It seems that you see a con-

[1] https://www.mhanational.org/issues/state-mental-health-america

Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 219 - 229

nection between your smoking and the possibility of having another heart attack.

Motivated by the importance of reflective listening skills and the significance of real-time feedback in the success of a counseling encounter, we envision our system as an automatic assistant that provides counselors with sample reflection language that is appropriate to the conversation context, thus helping counselors to acquire or improve reflective listening skills by emulating traditional psychotherapy training, but without the need of close human supervision.

We present a reflection generation system that leverages state-of-the-art language models, and further improve it with context augmentation techniques inspired by traditional counselor training. Specifically, we (1) identify previously used reflections from related sessions based on the current context, similar to how trainee counselors are exposed to several types of reflections on the same topic before they have to produce their own; and (2) we expand the content with synonyms for verbs and nouns, similar to how counselors are advised to use rephrasing strategies such as synonym rewording (Flasher and Fogle, 2012).

We perform a domain adaptation on an additional counseling corpus containing a variety of counseling styles, and fine-tune our system on a corpus of successful counseling interactions with labels available. Thus, it allows the system to benefit from successful counseling patterns derived from the cumulative experience of a large number of professionals. We conduct several comparative experiments, and perform evaluations using automatic metrics for language generation, including n-gram based, embedding-based and language diversity metrics. In addition, given the subjective nature of our task and the inability of automatic metrics to capture other relevant aspects of reflection generation, we conduct a human evaluation to assess the ability of our system to generate counseling reflections that are grammatically correct, fluent, and relevant to the conversation context.

2 Related Work

There have been significant efforts put in building automatic tools that provide support for mental and behavioral health. In particular, for dialogue-based counseling most of the existing work has focused on generating conversational agents that emulate the counselor in chat-bot like settings. For instance, (Han et al., 2013) built a system that extracts 5w1h (who, what, when, where, why, and how) information and user emotions (happy, afraid, sad, and angry) to recognize what the user says, predict the conversation context and generate suitable responses based on utterance templates developed to encode three basic counseling techniques (paraphrasing, asking open questions, and reflecting feelings). A similar system is presented in (Han et al., 2015), where authors first detect the user emotion and intention (e.g., greeting, self-disclosure, informing, questioning) and then extract the entities present in the utterance as well as related information (from an external knowledge base) to generate an appropriate response using language templates.

While these studies have focused on the delivery of health interventions via conversational agents (i.e., virtual counselors), we seek to build an automatic dialogue generation system that can help training counselors to improve their everyday practice. This is in line with a recent study on the impact of technology in psychotherapy, which has identified the development of technologies for counselor's training and feedback and technology-mediated treatment as important needs in this domain (Imel et al., 2017). Initial work in this direction is presented in (Tanana et al., 2019), where authors present a system that implements an artificial standardized client that interacts with the counselor and provides trainees with real-time feedback on their use of specific counseling skills by providing suggestions on the type of skills to use. Following the same line of work, our goal is to aid counselors while training specific skills, more specifically reflective listening skills. However, different from previous work, we focus on presenting the counselor with automatically generated samples for potential reflections that can be used immediately in the conversation.

Finally, potential applications of our proposed system include supporting counselor training in counseling platforms such as Talkspace[2], which currently has over a million users and five thousand therapists, and Crisis Text Line,[3] with 20 thousand counselors, handling over three thousand conversations a day, allowing users to connect with licensed therapists and to seek help via text messaging. The ability to automatically generate reflections given

[2]https://www.talkspace.com/
[3]https://www.crisistextline.org/

a conversation context can assist these counselors in formulating what they are going to say, thus improving the efficiency and quality of their reflections, with the final goal of increasing the number of people they can help and the effectiveness of their interaction on patient outcomes.

3 Model Overview

To build an automatic reflection generation system, we rely on the Generative Pretrained Transformer 2 (GPT-2) architecture (Radford et al., 2019) as a base model. GPT-2 is a state of the art transformer-based general purpose language model that has been found useful for dialogue generation tasks (Zhang et al., 2019). Our choice is motivated by its ability to produce language that closely emulates text written by humans (Wolf et al., 2019b).

Our model learns how to generate a counselor reflection using a GPT-2 architecture by operating entirely in a sequence-to-sequence way. In order to condition the generation on the counseling dialogue context and to generate reflections that are stylistically correct, we fine-tune the model with conversations in the counseling domain.

Below, we describe important elements of the model architecture related to the reflection generation task.

Input representation. The input sequence for the model consists of a counselor's utterance and a dialogue context including previous utterances from either the client or counselor. The window size of the dialogue context is set to five utterances, as a larger window size did not improve performance in preliminary experiments.

Embeddings. Besides learning word and positional embeddings, we also learn type embeddings to indicate whether the current token is part of the utterance from the client, counselor, or the reflection response. We use a trainable embedding matrix to map each location or type into a vector with the same size as the token embeddings. Separation tokens are also added to further delimit these elements in the dialogue.

Decoding details. The generator model consists of a transformer decoder with a similar structure to the decoder in (Vaswani et al., 2017) but only keeping the self-attention blocks. During the decoding stage, we assume we only have access to the augmented input and dialogue context and not the response. At each time-step, the model chooses a token from the output distribution conditioned on the context and the previously decoded tokens. The chosen token will be added into the input in the next time-step. To generate more diverse and expressive reflections, we adopted the top-k random sampling method (Holtzman et al., 2019), where the model samples from the k options with the highest probabilities.

4 Counseling-style Reflection Generation

Our goal is not only to generate natural-looking text that is relevant to the prompt but also to resemble the language style that counselors use while generating reflections. Thus, we extend the base model to incorporate two strategies that are commonly used by counselors while generating reflective statements.

First, we consider a training scenario where trainees are first shown sample reflections made while discussing different behavioral change goals (e.g. smoking cessation or weight management). After they have been exposed to several types of reflections, trainees are usually asked to construct alternative reflections for a given scenario as a way to reinforce what they have learned. In this case, trainees might associate previous reflections with the same behavioral change target as potential examples to generate their own. We attempt to use the same strategy to improve our system's responses. Thus, we devise a retrieval-based method to obtain a reflection to be used to expand the dialogue context.

Second, considering that counselors generate reflections using rephrasing strategies such as rewording with synonyms and verb tense changes, we design a content expansion method that augments the system input with verb and nouns synonyms. These methods are described in detail below.

4.1 Retrieval of the Most Similar Reflection

We seek to identify reflections that contain wording that could be useful for generating an appropriate reflection given the dialogue context. This is done in two main steps.

Selecting a relevant conversation. We start by identifying a set of relevant conversations i.e., conversations discussing the same behavior change. We then calculate the semantic similarity between the current dialogue context and this set of conversations. More specifically, we use TF-IDF (term frequency-inverse document frequency) encoding

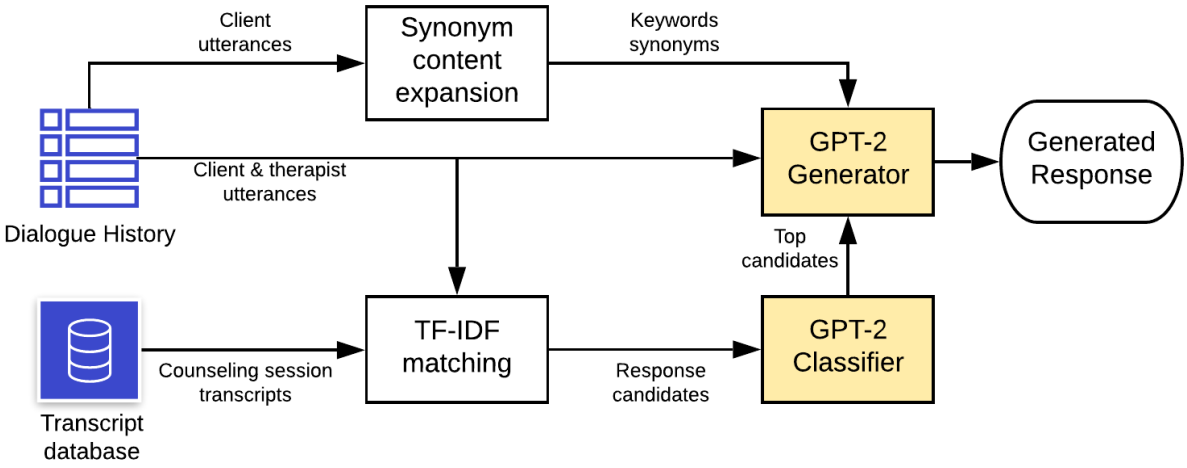

Figure 1: Model architecture. The fine-tuned model uses only client and therapist utterances, while the retrieval and content expansion models include additional input (TF-IDF matching and synonym content expansion) for the generation model.

Class	Precision	Recall	F1 score
In context	0.768	0.779	0.773
Not in context	0.765	0.754	0.759

Table 1: Performance metrics for the reflection-in-context classifier

for the dialogue context and candidate conversations and calculate their cosine similarity. We then select the conversation with the highest similarity as the most relevant conversation given the context. This stage may be further improved with methods such as BM25 or neural-based matching in future work.

Selecting a candidate reflection. Our next step focuses on identifying, among the reflections made in the most similar conversation, which of them is more likely to be a good match to the current context. The selected reflection is then added to the input of the generation system as a way to provide wording alternatives. For this task, we first build a set of candidate pairs by concatenating the current dialogue context and each of the reflections made in the most similar conversation. Then, we feed them to a binary classifier that aims to classify whether a sequence contains a valid reflection according to the given context. We score each sequence using the probabilities provided by the classifier and choose the one with the highest score as the best example reflection to be added to our current dialogue context.

To build the reflection-in-context classifier, we use a GPT-2 model and modify it by adding a classification layer to the output layer. The classifier is trained on a balanced set, with positive samples consisting of reflections from our main dataset, along with five previous utterances in the actual conversation, and negative samples consisting of reflections paired with random context windows taken from different conversations. We train the classifier using an 80%-20% split for training and testing sets respectively. The classifier achieves an accuracy of 76%, with detailed metrics per class shown in Table 1, thus showing reasonable performance on determining whether a reflection matches the current context.

4.2 Content Expansion

We augment the context content by applying synset expansion to synonyms and verbs. We first apply part-of-speech (POS) tagging on the context utterances using Stanford CoreNLP (Manning et al., 2014) to identify nouns and verbs and then obtain their corresponding synonyms for all their meanings using the English WordNet (Miller, 1998).

We then produce one rephrase for each utterance in the context by replacing the original nouns and/or verbs with a randomly selected synonym with the same POS tag. Our system uses the resulting utterances to augment the current context.

5 Experimental Setup

5.1 Counseling Datasets

We use the Motivational Interviewing (MI) counseling dataset from Pérez-Rosas et al. (2016) as the main corpus for training our retrieval and genera-

Total sessions	254
Vocabulary size	8,259
Total reflections	3,939
Average turns / session	97.2
Average tokens / reflection	20.9

Table 2: Statistics of the MI dataset

tion models, and perform language model domain adaptation using the Alexander Street dataset consisting of a variety of psychotherapy styles (e.g., cognitive behavioral, existential, solution focused). The datasets are described below.

MI Counseling Dataset: This dataset consists of 276 MI conversations annotated at utterance level with counselor verbal behaviors using the Motivational Interviewing Treatment Integrity 4.0 (MITI). In addition, the dataset also contains labels at the session-level, which evaluate the quality of the counseling interaction. The conversations portray MI encounters for three main behavior change goals: smoking cessation, medication adherence, and weight management. Among the different annotations available in the dataset, we focus on the annotations of counselor reflections, including simple reflections and complex reflections. Before we use the MI dataset, we remove transcripts corresponding to encounters that were deemed as low-quality counseling based on the global evaluation of the counseling interactions, i.e, sessions having low empathy scores or a low ratio of questions to reflections. We are thus left with a set of 254 counseling conversations. Dataset statistics are provided in Table 2. During our experiments using this dataset, we use 10% of the data as the test set and 5% as the validation set.

Alexander Street Dataset: This is a collection of psychotherapy videos that are published by Alexander Street Press.[4] The videos and its corresponding transcripts, containing psychotherapy conversations between clients and therapists on several behavioral and mental issues, are available through a library subscription. From this library, we downloaded the transcripts available under the Counseling & Therapy in Video: Volume IV, which contains around 400 real therapy sessions. However, due to the format inconsistencies, we were able to collect only 312 transcripts.

[4]http://alexanderstreet.com/

5.2 Reflection Generation Neural Architecture

During our experiments, we use a medium-size pre-trained GPT-2 (Radford et al., 2019) model as the backbone network for the language generation models. Our models are implemented using the Transformers library (Wolf et al., 2019a). The base model uses a byte-pair encoding (BPE) (Gage, 1994) and has a vocabulary size of 50,257. We use dropout with probability 0.1 for the embedding and attention layers and also for the residual connection in the blocks.

In addition, we use a warmup scheme for the learning rate using 5% of the total steps as warmup steps (Popel and Bojar, 2018). We use the Adam optimizer with weight decay (Kingma and Ba, 2015) to optimize the network at a learning rate of 6e-5. All models are trained for 10 epochs with early stopping.

5.3 Reflection Generation Experiments

We conduct two main sets of experiments on automatic reflection generation as described below. During our experiments we use the datasets described in section 5.1.

Reflection generation using a fine-tuned GPT-2 model. In this experiment we use the base model described in section 5.2 to generate counselor reflections. We first perform domain adaption of the language model using the Alexander Street dataset. We then fine-tune the generator using the MI dataset.

Reflection generation with retrieval and content expansion strategies. We extend the fine-tuned model to include the retrieval of the most similar reflection and content expansion strategies described in section 4.1 and 4.2. We experiment with incremental models that incorporate one strategy at the time.

Finally, we compare our models with a seq2seq model, which is frequently used as a baseline for conditional text generation problems (Vinyals and Le, 2015). We use the seq2seq implementation available in OpenNMT (Klein et al., 2017). The encoder and decoder are 2-layers GRU (Gated Recurrent Units) (Cho et al., 2014) with 512 hidden units. We train the model for 10 epochs with an Adam optimizer at a learning rate of 0.001.

Models	ROUGE			Embedding			Diversity		Avg Len
	RG-1	RG-2	RG-L	Greedy	Average	Extrema	Div-1	Div-2	
Seq2Seq	0.078	0.004	0.060	0.363	0.613	0.309	**0.156**	0.447	11.189
Fine-tuned GPT-2	0.152	0.020	0.117	0.446	0.726	0.382	0.134	0.496	18.522
+ retrieval	0.156	0.025	0.117	**0.456**	**0.735**	**0.390**	0.127	0.486	18.677
+ content expansion	**0.162**	**0.031**	**0.126**	0.453	0.731	0.386	0.128	**0.498**	18.412

Table 3: Performance of our models and the seq2seq baseline on the automatic generation of counselor reflections using ROUGE and embedding based metrics and n-gram diversity. We also show the average length of generated utterances for each model.

5.3.1 Automatic Evaluation Metrics

For the quantitative analysis of our reflection generation model, we use well-known automatic metrics for language generation, including:

ROUGE metrics: We use the ROUGE metric, a word overlap metric frequently used in the evaluation of neural language generation systems (Lin, 2004), including ROUGE-N, and ROUGE-L.

We decided to use ROUGE over other n-gram-based metrics, such as BLEU, because our task of generating reflective responses shares some similarity with the task of text summarization, where ROUGE is the metric of choice. Additionally, evaluations that we ran with other n-gram-based metrics had results consistent with those obtained with ROUGE.

Embedding-based metrics: We also use three embedding-based metrics, namely greedy matching, embedding average, and vector extrema (Liu et al., 2016). The first matches each token in one sentence to its nearest neighbor in the reference sentence, this metric favours generated reflections containing keywords that are semantically similar to the ground truth reflection. The other two calculate similarity for a pair of sentences based on their vector representations instead of matching each word. The sentence vector representations are constructed by averaging the word embeddings or taking the number with the highest absolute value for each dimension.

Diversity: We also evaluate diversity by measuring the ratio of distinct n-grams in the generated reflection with respect to the reference reflection.

5.3.2 Human Evaluation for Reflection Generation

To assess our automatic reflection generation systems' ability to produce relevant and coherent reflections, we also conducted a human evaluation study. We recruited two annotators familiar with counseling reflections, and asked them to evaluate the generated outputs and the ground truth responses for 50 samples randomly chosen from our test set. Given the conversation context of the latest five utterances, the annotators are asked to evaluate three main properties of several response candidates: relevance, reflection-likeness, and quality. The candidates are composed of the ground truth response and generated responses from four systems, i.e. seq2seq, GPT fine-tuned, and two improved versions using retrieval and content expansion. The annotators evaluate one candidate at a time, without knowledge of its origin.

Quality is evaluated using a 5-point Likert scale (i.e., 5: very good, 4: good, 3: acceptable, 2: poor and 1: very poor). We chose a 3-point Likert scale (i.e., 1: not at all, 2: somewhat, 3: very much) to evaluate relevance and reflection-likeness, since a finer scale may exceed the annotators' discriminating power (Jacoby and Matell, 1971). More specifically, we use the following prompts:

Relevance: Does the response seem appropriate to the conversation? Is the response on-topic?

Reflection-likeness: Does the response show understanding of the feelings of the client? Does the response paraphrase or summarize what the client has said?

Quality: How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?

We measured inter-rater agreement using Krippendorff's α (Krippendorff, 2018) and obtain agreement values of 0.18, 0.23, and 0.12 for relevance, reflection-likeness, and quality, respectively. The subjective nature of the question prompts may be the main reason for the low to fair levels of agreement on the different categories. The difference in

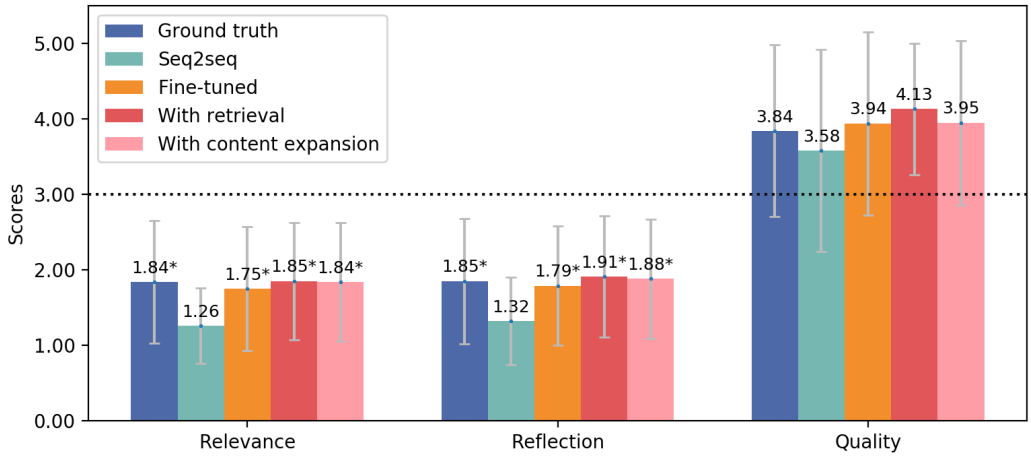

Figure 2: Human evaluation mean scores and standard deviations on the three criteria: relevance, reflection-likeness, and quality. (The former two criteria are in 3-point Likert scales. Quality uses a 5-point Likert scale; "*" indicate statistically significant improvement ($p<0.01$) over the seq2seq baseline)

personal preference and the level of background knowledge can both be sources of disagreement (Amidei et al., 2018). We plan to use more sophisticated evaluation schemes in future work, such as magnitude estimation or RankME (Novikova et al., 2018), instead of a plain Likert scale.

6 Results

6.1 Automatic Metrics

Table 3 reports scores for our models and the seq2seq baseline. From this table, we observe that all our proposed models outperform the seq2seq baseline as measured by the different metrics. In addition, our models with context augmentation (i.e., including retrieval of the most similar reflection and content expansion) outperform the fine-tuned model, thus suggesting that the proposed retrieval and expansion strategies are useful to improve the generation of reflections. Interestingly, the generation model augmented with the most similar reflection scores higher when using the embedding metrics, thus indicating that the model benefits from augmenting the context with words that are semantically close to it. Similarly, when using context expansion, we observe improved scores for the ROUGE-based metrics as the model takes advantage of the additional wording alternatives.

6.2 Human Evaluations

The average scores for each system response on *relevance*, *reflection-likeness* and *quality* are shown

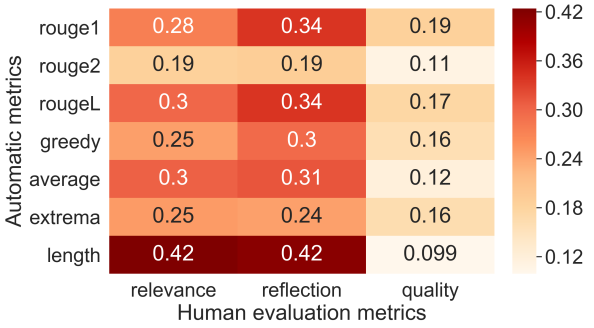

Figure 3: Spearman's correlation between human evaluation metrics and automatic metrics

in Figure2. From this figure, the general trend indicates that our systems perform on-par or above the reference reflections (ground truth), and outperform the baseline with statistical significance for both, relevance and reflection-likeness.

More specifically, in terms of *relevance*, we observe that our context-augmented models score the best by providing reflections that are more in line with the conversation context. Furthermore, they score very close to the ground truth reflection. In contrast, the seq2seq output obtains the lowest scores, thus suggesting that the generated reflections are more likely to be off-topic. The *reflection-likeness* follows a similar trend, confirming that the retrieval and expansion strategies do improve the model context.

For *quality* assessments, all systems (with the exception of the seq2seq model) produce responses considered to be of good quality according to the

Counseling Dialogue Context	Model	Generated Reflection
C: Um, yes, um I think, um, more so in terms of my family history, um and then suddenly last week I got the diagnosis of pre - diabetes. Um, and um, that's kinda thrown a wrench into my otherwise healthy normal you know life that I think I lead. **T:** So you are worried that this diagnosis is going to impact your day to day activities? **C:** Um, I'm worried a whole lot because my dad has, um diabetes and he uh, is right now going under dialysis,... So, the idea that I might be facing some of these issues **in my future** is horrendous for me.	Ground truth	Sounds like you really want to make some changes just to, as a prevention way to make sure you don't have to go through some of the things your farther went through.
	Seq2seq	Okay, so no missed meds.
	Fine-tuned	So you are really concerned about your weight.
	+ Content expansion	So you are really concerned about your health and where it might go **in the future**.
C: The joy of smoking! **T:** It's just, there's something that is enjoyable **C:** Yes! It's habit! I mean it is a habit. And there are certain things that make me smoke, you know, reasons I smoke every day **T:** Uh huh, **C:** It's the nicotine, is a piece of it. But it's not the only piece. **Retrieved:** So it's really becoming realistic in terms of how this habit that you've gotten used to can play a part in your other behaviors and your activities that you enjoy too.	Ground truth	Uh huh, ok, so you enjoy it.
	Seq2seq	In the trash can.
	Fine-tuned	There are other things that make you smoke, and that you like to have, so there's that, there's some withdrawal or other symptoms that make you smoke.
	+ Retrieval	There are other things that you enjoy doing, okay!

Table 4: **Top:** A case where content expansion helps. **Bottom:** A case where retrieval leads to an off-topic response. Examples of the generated reflections using the seq2seq baseline and our proposed systems. We show the dialogue context along with the ground truth reflection, where C stands for client, T stands for therapist. If the generated reflection consists of multiple utterances, we only show the first utterance.

Likert scale. Particularly, the system with retrieval scores 4.8% higher than the system with only fine-tuning. This can be attributed to the model conditioning the generation on sentences that have similar content but somewhat varying structure, thus allowing the model to take advantage of this information. The content expansion model performs similarly to the model with just fine-tuning. A potential explanation for this is that the content expansion sometimes produces subject-verb disagreement thus introducing noise during the generation process.

Finally, the *reflection-likeness* aspect of our evaluation obtains the highest scores from our retrieval model, followed by the content expansion model.

Human correlation analysis. To further validate our models, we conduct a correlation analysis between automatic metrics and human assessments as shown in Figure 3. In this analysis, we use Spearman's correlation because we care more about the monotonic relationship of the metrics instead of a linear relationship. From the results, we observe that the automatic metrics show weak positive correlations with human evaluations of *relevance* and *reflection-likeness*. Moreover, the *quality* evaluation shows a weak correlation with automatic metrics, which is somehow expected as n-gram-based metrics and embedding-based metrics do not take grammar into consideration. Similarly, the average length of generated reflections has almost no impact on whether the response is fluent or contains grammatical errors. On the other hand, average length obtains the highest correlations with reflection-likeness and relevance, suggesting that a longer reflection is more likely to contain information the client has previously mentioned.

6.3 Qualitative Analysis

To gain further insights into how the augmented input helps with generation, we analyze a sample output for our different systems as shown in Table 4. From this table, we observe that all models based on the pre-trained GPT-2 are able to generate reflections that agree, to some extent, with the dialogue context.

For the counseling conversation shown in the upper side of the table, we observe that the seq2seq model generates an off-topic reflection while the reflections generated by the other systems seem to be more relevant to the context. Therefore, showing the effectiveness of transfer learning for counseling-style reflection generation. More interestingly, when using content expansion the sys-

tem is able to generate a reflection with the phrase "in the future" as a more specific response, which further confirms that our expansion strategy does strengthen the signal of important information that we want the model to capture.

We also observe cases where our methods introduce noise in the reflection generation system. For example, in the counseling conversation shown in the bottom section of Table 4, the model trained without augmented context produces the most appropriate response. The retrieved sentence successfully captures the idea of "habits," while the conversation is about reasons other than habits that make the client to enjoy smoking, thus leading to the generation of a less relevant reflection.

7 Conclusion

We presented a system based on a state of the art language model that generates counseling reflections based on the counselor-client dialogue context. We first conducted domain adaptation and subsequently fine-tuned the system with motivational interviewing conversations. We then improved the system by augmenting the dialogue context using retrieval and content expansion methods that implement actual strategies used by counselors while generating reflections.

We conducted comparative experiments between systems implementing these strategies and demonstrated their effectiveness in generating improved reflections as measured by standard language generation metrics such as ROUGE as well as embedding-based and diversity metrics. To further validate our models, we conducted a human evaluation study on the generated responses. The evaluation showed that humans scored our proposed systems higher than the baseline model on quality, relevance, and reflection-likeness.

We believe that counselors could benefit from the proposed system by using the automatically generated reflections as reference while learning to formulate reflective statements.

Acknowledgements

We are grateful to Christy Li, Yinwei Dai, Jiajun Bao, and Allison Lahnala for assisting us with the human evaluations. 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

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Lydia V Flasher and Paul T Fogle. 2012. *Counseling skills for speech-language pathologists and audiologists*. Cengage Learning.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133, Prague, Czech Republic. Association for Computational Linguistics.

Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. 2013. Counseling dialog system with 5W1H extraction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 349–353, Metz, France. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Zac E Imel, Derek D Caperton, Michael Tanana, and David C Atkins. 2017. Technology-enhanced human interaction in psychotherapy. *Journal of counseling psychology*, 64(4):385.

Jacob Jacoby and Michael S Matell. 1971. Three-point likert scales are good enough.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

William R Miller and Stephen Rollnick. 2013. *Motivational interviewing: Helping people change, Third edition*. The Guilford Press.

William R Miller, Carolina E Yahne, Theresa B Moyers, James Martinez, and Matthew Pirritano. 2004. A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of consulting and Clinical Psychology*, 72(6):1050.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019. Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1448–1459, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *J Med Internet Res*, 21(7):e12529.

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.