A Comparative Multidimensional Analysis of Empathetic Systems

Andrew Lee[†] **Jonathan K. Kummerfeld**[†‡] **Larry An**[†] **Rada Mihalcea**[†]
[†] University of Michigan [‡] University of Sydney
ajyl@umich.edu

Abstract

Recently, empathetic dialogue systems have received significant attention. While some researchers have noted limitations, e.g., that these systems tend to generate generic utterances, no study has systematically verified these issues. We survey 21 systems, asking what progress has been made on the task. We observe multiple limitations of current evaluation procedures. Most critically, studies tend to rely on a single non-reproducible empathy score, which inadequately reflects the multidimensional nature of empathy. To better understand the differences between systems, we comprehensively analyze each system with automated methods that are grounded in a variety of aspects of empathy. We find that recent systems lack three important aspects of empathy: specificity, reflection levels, and diversity. Based on our results, we discuss problematic behaviors that may have gone undetected in prior evaluations, and offer guidance for developing future systems.[1]

1 Introduction

Empathetic dialogue systems have received significant attention in recent years, with new models that incorporate emotion, common sense, knowledge graphs or other signals into language models (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Kim et al., 2021). Meanwhile some researchers have noted that recent systems tend to generate generic, trite responses (Wang et al., 2022b; Sabour et al., 2021). However, these observations have not been verified in a systematic way.

We survey 21 empathetic dialogue systems, using new analysis methods to see what progress has been made. Quantitatively comparing these systems pose multiple challenges. Automated metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), or ROUGE (Lin, 2004) compare the *lexical overlap* between the generated text and a "ground-truth" discourse. However, in dialogue, there is an unbounded space of valid responses that differ from a ground-truth sample, and researchers have shown that these metrics have only a weak correlation with human judgement (Liu et al., 2016).

To mitigate this issue, recent empathetic systems often rely on human evaluations. These crowdsourced evaluations typically measure empathy level, fluency, and relevance, with the latter two measuring overall conversational quality, with no relation to empathy. Using only one measure of empathy is an overly simplified assessment of an empathetic dialogue system (Lahnala et al., 2022), failing to account for how empathy is a multidimensional construct (Davis et al., 1980a; Davis, 1983) with a wide range of definitions in terms of social, emotional, or cognitive dimensions (Cuff et al., 2016). One issue with using a single score is that some systems might be effective in one aspect of empathy, while other systems might excel in others, but that variation would be hidden when considering a single overall empathy score. Another issue is that a single score does not provide information about the nature of remaining errors, which would be valuable for guiding future work.

Given these limitations of prior evaluation procedures, we comprehensively study nine systems on multiple metrics that are each grounded in a multidimensional definition of empathy. Namely, we survey recent papers that propose an empathetic dialogue system from NLP conferences in the last three years. Of the 21 systems that we identify, we further analyze every system that has been trained on EmpatheticDialogue (Rashkin et al., 2019), the most prominent dataset used by researchers (Table 1), and has been open-sourced, which results in nine systems. Through our study we find that recent systems lack specificity, reflection levels (Houck

[1]The data and scripts we used in this comparative analysis can be found at https://github.com/MichiganNLP/empathy_eval

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 179–189

et al., 2012), and diversity, each of which may have gone undetected with prior evaluation procedures.

Our study provides a reflection on the advances made by recent empathetic systems, and offers valuable takeaways for the development of future systems.

2 Related Work

Empathetic Dialogue Systems. The majority of recent empathetic dialogue systems train language models with examples of empathetic responses. EmpatheticDialogues (Rashkin et al., 2019) has become a popular choice for such data, consisting of 25k conversations grounded in emotional situations. Researchers often additionally incorporate sentiment or emotion (Majumder et al., 2020; Rashkin et al., 2019; Lin et al., 2019), common sense (Sahand Sabour, 2021), or knowledge (Li et al., 2022) in order to ground the conversations to real life human experiences.

Empathy Frameworks. Empathy is a nuanced human experience and a complex multidimensional construct which is difficult to computationally assess. Broadly speaking, empathy has two aspects: emotion and cognition (Davis et al., 1980b). The emotional aspect relates to the emotional reaction or connection that is formed as a reaction to one's emotions or experiences, whereas the cognitive aspect relates to the reflective and interpretive process of understanding one's experiences.

Definitions from psychotherapists provide the foundations for computational researchers to assess the empathy level of their systems. For instance, EPITOME (Sharma et al., 2020) identifies empathy across three "Communication Mechanisms" (emotional reactions, interpretations, exploration). Liu et al. (2021b) grounds their emotional support conversation framework on Hill's Helping Skills Theory (Hill, 2009), consisting of three stages of support (exploration, insight, action).

However, by surveying recent empathetic systems, Lahnala et al. (2022) critically indicate that most systems lack a clear definition of empathy. Our findings confirm that of Lahnala et al. (2022), and go beyond their work by providing empirical studies.

3 Limitations of Current Evaluations

Recent empathetic dialogue systems follow a common procedure to evaluate their output, which includes automated metrics and human evaluations.

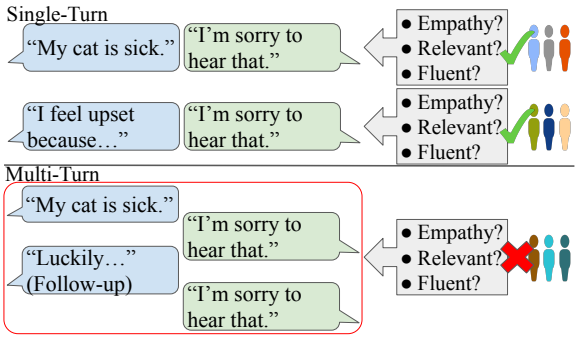

Figure 1: Evaluating on samples of single-turn evaluations, distributed to multiple judges, can appear empathetic, fluent, and relevant, despite issues such as repetition.

However, they have several shortcomings, which we discuss in this section. Namely, we survey the evaluation procedure of 21 empathetic dialogue systems published over the last five years at leading NLP venues, as summarized in Table 1. The following sections describe the metrics and evaluation procedures followed by these systems, along with their limitations.

3.1 Human Evaluations

Given the nuanced construct of empathy, every surveyed system includes human studies, which typically consist of: Likert-scale questions to measure empathy, fluency, and relevance of generated responses, and an A/B test to compare preferences between empathetic systems: Both of these are typically done with the help of crowdworkers.

Likert-Scale Questions. Current approaches for crowdsourcing a single empathy score have two key limitations. First, given the multidimensional nature of empathy, it is difficult to assign a single empathy score that captures the various aspects of empathy. This makes it difficult to attribute system behavior to empathy, as some systems might be effective in a specific aspect while other systems might excel in others. By the same token, a single score makes it difficult to understand how to improve each system.

Second, current evaluations are typically conducted on samples of single-turn exchanges (ie, a single pair of a prompt and a response). To understand why this is a major limitation, consider a simple *IF_ELSE* system that always generates either *"I'm happy to hear that."* or *"I'm sorry to hear that."* based on the sentiment of the input utterance. While such a system is not meaningfully empa-

System	Dataset	Automated Metrics	Human Evaluation
EmpatheticDialogue (Rashkin et al., 2019)	ED	PPL, BLEU	Likert Scale, A/B Test
MoEL (Lin et al., 2019)	ED	BLEU	Likert Scale, A/B Test
MIME (Majumder et al., 2020)	ED	BLEU	Likert Scale, A/B Test
EmoCause (Kim et al., 2021)	ED	Coverage, Empathy Classifiers	Likert Scale, A/B Test
Dual-Emp (Shen et al., 2021)	ED	PPL, BLEU, Dist-n, Embed	Likert Scale, A/B Test
Gao et al. (2021)	ED	BLEU, Dist-n, Embed	Likert Scale, A/B Test
KEMP (Li et al., 2022)	ED	PPL, Dist-n	Likert Scale, A/B Test
CEM (Sabour et al., 2021)	ED	PPL, Dist-n	A/B Test
EmpHi (Chen et al., 2022)	ED	BLEU, Dist-n	Likert Scale, A/B Test
Emp-RFT (Kim et al., 2022)	ED	PPL, Dist-n, BERTscore	Likert Scale, A/B Test
CARE (Wang et al., 2022a)	ED	PPL, BLEU, BERTScore	Likert Scale, A/B Test
SEEK (Wang et al., 2022b)	ED	PPL, Dist-n, DE, UEI, REI	Likert Scale, A/B Test
Lee et al. (2022)	ED	Empathy Classifiers, dist-n, NIDF, PPL	Likert Scale, A/B Test
EMOTICONS (Colombo et al., 2019)	Cornell, OpenSubtitles	BLEU, Dist-n	AffectButton
CoBERT (Zhong et al., 2020)	PEC	R@n, MRR	None
CoMAE (Zheng et al., 2021)	Reddit	PPL, BLEU, ROUGE, Embed	Likert Scale, A/B Test
ESC (Liu et al., 2021b)	ECS	PPL, BLEU, ROUGE, Embed	A/B Test**
Liu et al. (2021a)	MojiTalk	PPL, Embed, TTR-n, Avg. Len, % Stopwords	Likert Scale
EDOS (Welivita et al., 2021)	EDOS	PPL, Dist-n, Embed	None
Zhu et al., 2022	MPED	ROUGE, BLEU	Likert Scale, A/B Test
Cheng et al. (2022)	ESC	PPL, BLEU, ROUGE, METEOR, CIDEr	A/B Test

Table 1: Recent empathetic dialogue systems from NLP conferences and their evaluation methodologies.

thetic, when a single sample is distributed across multiple evaluators, its responses will always be considered empathetic, fluent, and relevant, and its repetitive behavior goes undetected (Figure 1). As it turns out, many recent empathetic dialogue systems indeed repeat the same responses for different input utterances, which we analyze in Section 5.4.

A/B Tests. For A/B tests, single-turn dialogues from two different systems are sampled, and crowdworkers are asked to select the system that they prefer. Such an approach suffers from similar issues. Namely, when system A is preferred over B, it is unclear how to interpret the preference in terms of aspects of system behavior. For instance, system A might tend to convey emotional empathy while system B conveys cognitive empathy. Furthermore, pairwise comparisons amongst systems is unscalable and rather cumbersome.

3.2 Automated Metrics

Because human evaluation is expensive, researchers often include automated metrics. The most commonly used automated metrics include BLEU, ROUGE, BERTScore (Zhang et al., 2020), PPL, and Distinct-n (Li et al., 2016).

BLEU, ROUGE, BERTScore. These metrics compare generated responses against a known ground-truth utterance. BLEU and ROUGE use lexical overlap, while model-based approaches such as BERTScore use similarity scores in high dimensional spaces. While these may be suitable metrics for tasks such as translation or summarisation, dialogues often have an unbounded number of valid responses that all differ semantically. Given the open-ended nature of dialogue, comparing system responses against ground-truth utterances is misleading, and Liu et al. (2016) demonstrate that BLEU and ROUGE scores share little correlation with human judgement.

Perplexity. Perplexity (PPL) measures the degree of uncertainty of a language model in the sequences it generates, and while it is a useful intrinsic evaluation of a language model, it does not necessarily characterize the behavior of a model on a specific

task (ie., empathetic response generation).

Distinct-n. Distinct-n is a measure of diversity, calculated by dividing the number of distinct n-grams generated by the total number of generated tokens. While this metric captures the variance in token distributions of predicted responses, it is difficult to interpret these values.

Furthermore, current measures of diversity do not distinguish utterance-level and turn-level diversity. We find that recent empathetic dialogue systems often repeat the same phrase for multiple prompts. These behaviors are not properly reflected with current measure of diversity.

Empathy Detection. Lastly, some studies suggest the use of automatic empathy detection models such as EPITOME (Sharma et al., 2020) to evaluate the empathy level of dialogue systems (Lee et al., 2022; Kim et al., 2021). However, Lee et al. (2023) demonstrate that EPITOME does not always use dialogue context, but rather rely on phrases such as "I'm sorry to hear that."

4 Experimental Setup

4.1 Analyzing Multi-Turn Dialogues

Given the multidimensional and personal nature of empathy, we do not view "state-of-the-art" on a single empathy score as a useful measure. Rather, we compare model behavior against that of people, with respect to metrics grounded in various aspects of empathy.

Most prior studies have only evaluated systems on single-turns. However, such an evaluation can overlook specific model behaviors, such as repetition (Section 3). In our experiments, we analyze the *multi-turn* behavior of systems.

Evaluating multi-turn behavior requires that we generate multi-turn conversations. However, in order to compare across systems, the prompts must be controlled. Namely, given a dialogue for a human (H) in EmpatheticDialogue as a sequence of prompts and responses $(P_0, R_0^H, ..., P_n, R_n^H)$, a meaningful comparison for system S would be a sequence with the same prompts $(P_0, R_0^S, ..., P_n, R_n^S)$. However, there is no easy way to generate such a sequence, as all subsequent prompts P_i depend on the previous context $(P_j, R_j \text{for all } j < i)$.

Thus, rather than constructing multi-turn sequences for each system, we measure multi-turn metrics in a *piece-wise* manner, by providing segments of the dialogue context incrementally. That is, we deconstruct each human dialogue consisting of n-turns $(P_0, R_0^H, ..., P_{n-1}, R_{n-1}^H)$ into n contexts, and use each one as input to our system:

$$\forall i < n, R_i^S = S.generate(P_{0:i-1}, R_{0:i-1}^H)$$

Although the resulting final sequence is likely incoherent (ie, R_i^S followed by P_{i+1} may be incoherent), each utterance is valid in the provided context. Given that, we calculate metrics at each point and simply aggregate the mean metric values for the generated responses $R_{0:n}$. We use our piece-wise multi-turn evaluation setup for all metrics described in Section 5.

Note the subtle difference between our *piece-wise* multi-turn evaluation and an *interactive* multi-turn evaluation – in our setting, a human evaluator is not interacting and evaluating at every turn.

4.2 Surveyed Systems

Of the 21 systems in Table 1, we analyze every system that is trained on EmpatheticDialogue and is open-sourced, resulting in nine systems. We only consider systems trained on EmpatheticDialogue, as it is the most widely used dataset (Table 1), but also in order to control for the data that each system is trained on. We consider two baselines: the original system proposed by the authors of EmpatheticDialogue, and human responses, which are the human utterances in the test split of EmpatheticDialogue.

4.3 Data

We use EmpatheticDialogue (Rashkin et al., 2019) for our experiments, which consists of 25k crowdsourced conversations. Each multi-turn dialogue is constructed from a pair of workers, in which the first worker is instructed to describe a situation in which they have experienced a specific emotion. The two workers then have a conversation around the experience. The data consists of an official train, validation, and test set from a 8:1:1 split. After verifying that each system that we survey uses the same data splits, our experiments are conducted on the test split, consisting of 2547 conversations, or 5255 turns (where each turn consists of two utterances, one from each party).

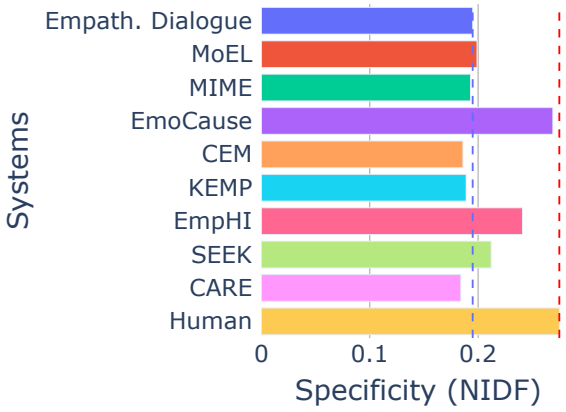

Figure 2: Specificity scores (NIDF). Blue and red lines indicate baselines from Empathetic Dialogue and human, respectively.

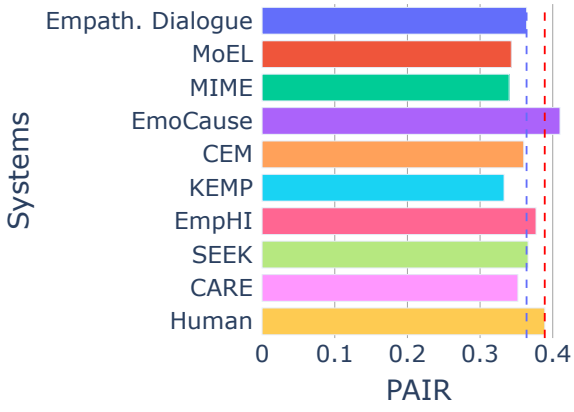

Figure 3: PAIR scores from each system. Blue and red lines indicate baselines from Empathetic Dialogue and human, respectively.

5 Multidimensional Evaluations of Empathy

Prior work from psychology that studies effective ways of evaluating empathy in counseling settings has defined various aspects of empathy (Truax and Carkhuff, 1964; Banerjee and Lavie, 2005; Elliott et al., 2005; Truax and Carkhuff, 2007; Houck et al., 2012). In this work, we focus on four specific aspects from Truax and Carkhuff (1964), Elliott et al. (2005), and Houck et al. (2012), which we discuss below. For each aspect, we provide our motivation for measuring the aspect, our methodology for measurements, and our results.

5.1 Specificity

Motivation. Truax and Carkhuff (1964) define concreteness, or specificity, as the degree to which a practitioner comments on generalities or abstract ideas (low specificity) versus specific feelings or experiences (high specificity). Specificity has a few benefits. First, it ensures that the practitioner's responses does not become abstract or emotionally removed from the patient's feelings and experiences. Secondly, it allows the practitioner to be more precise in understanding the client's feelings and experiences. Lastly, it encourages the client to attend closer to their problem areas or emotional conflicts.

Methodology. See et al. (2019) propose Normalized Inverse Document Frequency (NIDF) to measure the specificity of a dialogue. We use the same formulation:

$$NIDF(w) = \frac{IDF(w) - min\_idf}{max\_idf - min\_idf},$$
$$IDF(w) = log(R/c_w)$$

Where R is the number of samples in a dataset, c_w is the number of samples that contain w, and min_idf, max_idf are the minimum and maximum IDFs taken over all words in the vocabulary. The specificity score for a response r is the mean NIDF of the words in r.

System Evaluations. Figure 2 shows the NIDF scores for various systems. When compared against EmpatheticDialogue, we see that most systems have a very close similarity score (<= 0.005 difference), with four systems actually having lower scores. When compared against human behavior, we observe that all systems are less specific. The low and converged specificity scores may be related to the repetitive behavior shown in later sections (Section 5.4). These scores indicate the need to better understand whether systems appear empathetic because they tend to utter trite and generic phrases (1), or are offering concrete responses that demonstrate a relatable experience or emotion.

5.2 Reflection Level

Motivation. Houck et al. (2012) discuss the importance of *reflection* for therapists in conveying empathy. Reflection is one's ability to understand and reflect on what the client is saying, and is typically classified as simple reflection or complex reflection, with the latter being the preferred level of reflection to practice.

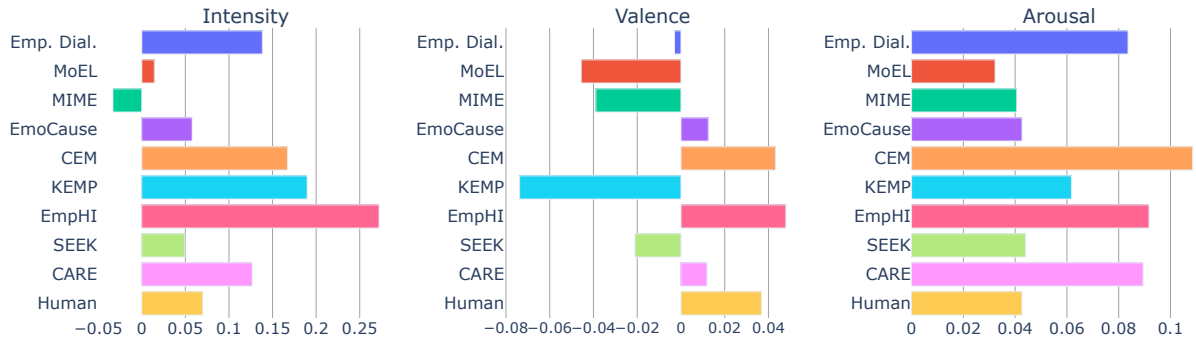

Figure 4: Difference in emotion intensity, valence, and arousal between prompt utterances and response utterances. A negative score indicates that the response had a greater intensity/V&A value than the prompts (only prompt utterances).

Methodology. We use PAIR (Min et al., 2022), a RoBERTa variant model that measures the reflection level of utterances in order to provide feedback for counseling trainees. Specifically, PAIR uses contrastive learning to rank an utterance as non-reflective, simple reflective, or complex reflective. PAIR outputs a continuous score between 0 and 1.

System Evaluations. Figure 3 demonstrates the PAIR scores of our surveyed systems. Compared to EmpatheticDialogue, 6 systems score lower, while compared to that of human behavior, all systems except for one score lower. Offering complex reflections is a challenge, in which even human responses offer low reflection scores. This suggests that researchers may need alternative approaches to build systems with complex reflections rather than immitating that of human behavior (Min et al., 2022; Sharma et al., 2021).

5.3 Word Choice

Motivation. When evaluating empathy, Elliott et al. (2005) demonstrate the correlation between the use of rich, vivid, and metaphorical language that is *consistent with the client's discourse* and the clients' perceptions of empathy. This component is sometimes referred to as high vs. low energy (Cochrane, 1974).

Methodology. To measure the level of rich and vivid language, we measure affect intensities, valence, and arousal.

Affects (emotions, feelings, attitudes) have varying degrees of intensity (eg., outrage versus irritated). We measure the emotional intensity of system responses using the NRC Emotion Intensity Lexicon (NRC-EIL) (Mohammad, 2018b), which consists of an intensity score between 0 to 1 for 10,000 terms associated with emotions.

Valence and arousal (V&A) are orthogonal measures of emotional states: valence is a measure of how pleasant or unpleasant one feels, while arousal is a measure of how energized or soporific one feels. Note that arousal is different from intensity – for instance, grief or depression can be low arousal but intense feelings. To measure V&A, we use the NRC-VAD Lexicon Mohammad (2018a), which consists of more than 20,000 words and their V&A scores, each ranging from 0 to 1.

To measure whether the choice of words and degree of energy in a response is *consistent* with that of the client's discourse, we measure the *difference* in affect intensity and V&A scores of the prompts and responses. The affect intensity and V&A scores of utterances are assigned by taking the maximum intensity or V&A scores of tokens in the utterances. Note that with our metric, a value closer to zero is desired.

System Evaluations. Figure 4 shows the difference in affect intensity and V&A scores between prompts and responses of each system. Because we are measuring differences, a value closer to zero is desired, while a negative value indicates that the response had a higher intensity or V&A score than the prompts. For many cases, we observe that earlier systems (MoEL, MIME) actually have better scores than newer systems, suggesting that aspects such as emotion intensity or arousal are being overlooked by current systems.

5.4 Diversity

Motivation. Lastly, diversity is a key attribute of human dialogue. While a repetitive system such as the previously mentioned *IF_ELSE* system might always appear as empathetic, relevant, and fluent based on single samples, it can hardly be consid-

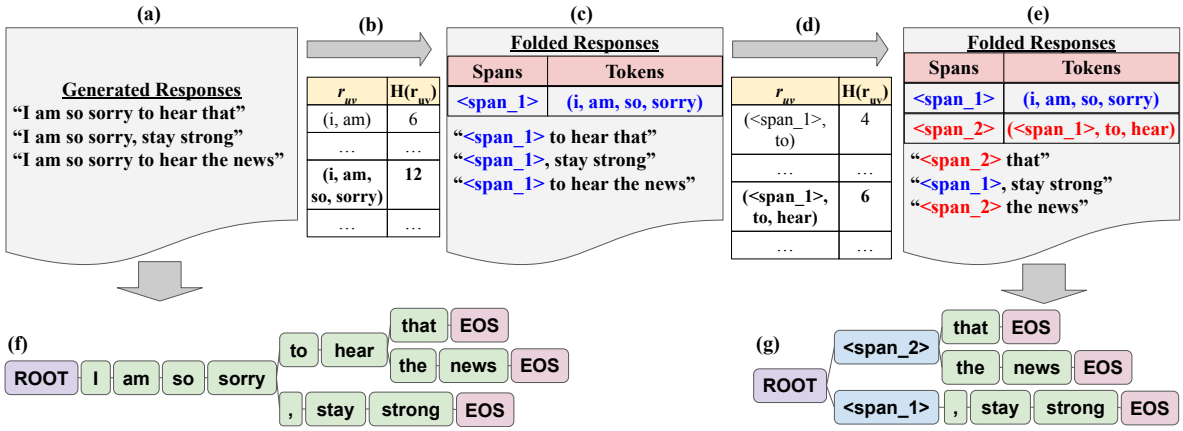

Figure 5: Folding procedure of our response-trie.

ered a meaningful system. For this reason we analyze the diversity of current system responses.

Methodology. Given the limitations of distinct-n (Section 3.2), we introduce a new measure of diversity, described in this section.

Response-Trie. Our diversity metric relies on constructing a *response-trie*, which we briefly describe here with details in subsequent sections.

The use of our response-trie is motivated by an observation that current systems frequently generate common sequences (ie, "I'm so sorry to hear that."). Given a set of generated responses, our procedure iteratively runs "folding" operations, which identify frequent sequences using a heuristic and replaces them with unique placeholder tokens. During our folding procedure, we maintain a mapping between placeholder tokens and the sequence that it has replaced. By the end of a series of folds, the original response set is converted to a set of "templates" that each system generated. By constructing a trie with the resulting templates, we can derive multiple metrics using various properties of our trie. An example of responses before and after our procedure is demonstrated in Figure 5 (a, e), and we describe our procedure in detail below.

Folding Responses. We notate a dialogue discourse from system S as (P_S, R_S), where P_S and R_S are each a set of prompts and responses. We notate each prompt $P_i \in P_S$ and response $R_i \in R_S$ as $P_i : p_{i0}, ..., p_{in}$ and $R_i : r_{i0}, ..., r_{im}$, where p and r are tokens in P and R.

The first step in constructing a response-trie is to iteratively "fold" the utterances in the response set R_S. The folding operation F defines a simple heuristic H for identifying spans, or n-grams, to fold. We use the product of the length of the span and the frequency in which it appears. That is, for each sequence of response tokens $r_{uv} = (r_u, ..., r_v) \in R_S$, $H(r_{uv}) = r_{uv}.length * r_{uv}.count$. Figure 5 (a, b) demonstrate an example of a response set R_S and H values from a single fold.

Once we identify the span r_{uv} with the highest heuristic value, we replace every occurrence of the span in R_S with a unique placeholder token (i.e., "<span_1>"), while maintaining a mapping between placeholder tokens and their corresponding spans (Figure 5, c). This folding procedure is repeated, while treating the newly inserted placeholder tokens like any other token, until a stopping criteria is met. We stop our folding procedure once none of the n-grams occurs more than once.[2]

After our folding procedure, we are left with a mapping between placeholder tokens and their corresponding n-grams, as well as a modified set of response utterances R'_S, where common n-grams in R_S are replaced by placeholder tokens (Figure 5, e). We refer to our modified utterances R'_S as *templates*. Each template $R'_i \in R'_S$ can be converted back to its original form R_i by substituting each placeholder token in R'_i according to our mapping.

Constructing a Response-Trie. Once our set of generated responses R_S has been iteratively folded into templates R'_S, we construct a trie T using R'_S (Figure 5, f, g).

Every token $r_i \in R'_S$, is represented as a node

[2]Note that we do not consider unigrams, as replacing a unigram with a placeholder would have no effect.

System	# Templates↑	# of Span Nodes / Total # of Nodes↓	# of Children (From Root)↑	Compression Ratio↑	# of Unique Start Words↑
Human Response	5201 (99.0%)	4974 / 37945 (13.1%)	1682 (32.01%)	60.35%	407
EmpatheticDialogue	2614 (49.7%)	1745 / 4434 (39.4%)	1091 (20.8%)	31.3%	37
MoEL	984 (18.7%)	832 / 1550 (53.7%)	588 (11.2%)	3.9%	20
MIME	4719 (89.8%)	3361 / 12573 (26.73%)	1322 (25.2%)	34.9%	17
EmoCause	4795 (91.2%)	3742 / 16453 (22.7%)	1950 (37.1%)	30.68%	759
CEM	795 (15.1%)	647 / 1241 (52.1%)	479 (9.1%)	38.6%	23
KEMP	925 (17.6%)	726 / 1456 (49.9%)	512 (9.7%)	35.9%	10
EmpHI	3386 (64.4%)	2289 / 6484 (35.3%)	1333 (25.4%)	34.95%	25
SEEK	1009 (19.2%)	888 / 1731 (51.3%)	634 (12.1%)	42.3%	23
CARE	1921 (36.6%)	1510 / 3217 (46.9%)	934 (17.8%)	33.5%	21

Table 2: Diversity metrics derived from our response-trie, based on 5,255 prompts from EmpatheticDialogue. Number of templates refers to the number of unique responses generated by each system. Compression ratio refers to the ratio between the size of tries made before and after our folding operations.

n_i in T. Directed edges in T preserve the order in which tokens occur. Namely, for a token sequence (..., r_{i-1}, r_i, r_{i+1}, ...), a directed edge exists from node n_{i-1} to n_i, and from n_i to n_{i+1}. We also include two special nodes $ROOT$ and EOS: every beginning token is connected to $ROOT$ and the last token of each response is connected with EOS. When adding a node n to T, if a path from the root node to n already exists, it is not added again. One can also imagine a set of responses R as a set of paths in T, in which each sub-path in the tree $n_{j:k}$ indicates a span of tokens that occur in R_S.

Metrics from Response Tries. Once the entire response body is encoded as a response-trie, we can use properties of the trie as metrics of diversity and gather qualitative insights. Examples include the number of nodes, number of children from the root node, or the compression rate of the tries before and after our folding operations (# of nodes after folding / # of nodes before folding), in which lower values imply more repetitive spans. Such properties from the response-trie provide qualitative insights that metrics such as distinct-n fail to provide.

System Evaluations. Table 2 shows a set of metrics derived from our response-trie. Most notably, we find that many systems are repetitive. This conclusion can be drawn by a few metrics. Of the 5,255 prompts given to each system, many systems have a much smaller number of templates that they generate. Many systems also have a high ratio of span nodes, indicating that a large portion of their responses consists of common phrases. Note that such system-wide repetitive behavior goes undetected in current evaluation methods when single samples are distributed to multiple human evaluators (Section 3.1).

6 Lessons Learned

Based on our evaluations, we formulate a set of takeaways, as well as concrete suggestions for evaluating future systems.

Single vs. Multidimensional Empathy. Single dimensional evaluations of empathy are not reflective of system improvements. Unlike a single empathy score, our analyses allow us to attribute system behavior to specific aspects of empathy. Overall, our takeaways echo the conclusions of prior work (Lahnala et al., 2022) that recent systems rely on an overly simplified, single dimensional definition of empathy, and highlight the shortcomings of current evaluation methods (Section 3).

Ablation Studies. On a similar note, systems need to be better ablated. While recent systems propose to incorporate emotion, common sense, or knowledge, the benefits of such additions are not being evaluated. Rather, broad strokes using a single empathy score are used to argue for improvement, which makes it difficult to tease apart the benefits of each suggested methodology.

Single vs. Multi-turn Interactions. Single-turn and multi-turn behavior of systems can portray vastly different pictures. For instance, problematic behaviors such as repetitions can go unnoticed when only considering single-turn samples. We argue that systems need to be evaluated on multi-turn interactions rather than single-turn samples.

Opensourcing Human Evaluations. Given the non-reproducible nature of crowdsourcing, we encourage researchers to openly share their crowd-

sourced results. Future work might ask crowdworkers to provide their reasons for their answers, which may better allow researchers to attribute system behavior to various aspects of empathy.

7 Conclusion

In this work, we surveyed recent empathetic dialogue systems. We discussed the shortcomings of the evaluation of these systems – relying on a single empathy score fails to capture the multidimensional aspect of empathy. By deploying several automated metrics, each grounded in different aspects of empathy, we identified multiple areas in which current systems could be improved, and uncovered behaviors that have gone undetected with previous evaluations, such as the lack of specificity, reflection levels, and diversity. Furthermore, we found that newer systems do not necessarily lead to improved performance under our metrics. We highlight the challenges of evaluating empathetic systems, and propose possible approaches to measure meaningful progress on the task.

8 Limitations

We acknowledge that evaluating empathy is difficult, and that a survey of recent systems is different from a proposal for future evaluations. That is, while our survey methodology may be suitable to discuss and uncover various limitations of current systems and their evaluation procedures, we do not show its suitability for future evaluations. In order for such metrics (or future metrics) to be suitable, we believe a human study is necessary in order to be used as a benchmarking tool.

Blindly relying only on automated metrics in applications of such systems, especially in a sensitive domain like healthcare and the mental health domain, can carry risks as well. Rather we encourage thorough examinations from practitioners, or that such systems be applied with humans in the loop.

Lastly, there is room for improvement for our evaluation of reflection levels using PAIR because of the possibility of a distribution shift from what PAIR was originally trained on vs. the empathetic dialogue that we are evaluating.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This project was partially funded by a National Science Foundation award (#2306372). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

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.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. EmpHi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074, Seattle, United States. Association for Computational Linguistics.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carolyn T Cochrane. 1974. Development of a measure of empathic communication. *Psychotherapy: Theory, Research & Practice*, 11(1):41.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. 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 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin M. P. Cuff, Sarah D. Brown, Laura K. Taylor, and Douglas Howat. 2016. Empathy: A review of the concept. *Emotion Review*, 8:144 – 153.

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44:113–126.

Mark H. Davis, Miles P. Davis, M Davis, Matthew Davis, Mark Davis, Mm Davis, M Davis, F. Caroline Davis, Heather A Davis, and Ilus W. Davis. 1980a. A multidimensional approach to individual differences in empathy.

Mark H Davis et al. 1980b. A multidimensional approach to individual differences in empathy. *Catalog of Selected Documents in Psychology*.

Robert Elliott, H F Filipovich, L. Harrigan, J. W. Gaynor, Cora Reimschuessel, and Judith K. Zapadka. 2005. Measuring response empathy : The development of a multicomponent rating scale.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Jon M. Houck, Theresa B. Moyers, William R. Miller, Lisa H. Glynn, and Kevin A Hallgren. 2012. Motivational interviewing skill code (misc) 2.5.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wongyu Kim, Youbin Ahn, Donghyun Kim, and Kyong-Ho Lee. 2022. Emp-RFT: Empathetic response generation via recognizing feature transitions between utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4118–4128, Seattle, United States. Association for Computational Linguistics.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew Lee, Jonathan Kummerfeld, Larry An, and Rada Mihalcea. 2023. Empathy identification systems are not accurately accounting for context. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1686–1695, Dubrovnik, Croatia. Association for Computational Linguistics.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does GPT-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683, Gyeongju, Republic of Korea. International Committee on 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.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

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.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. 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 121–132, Hong Kong, China. 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.

Ruibo Liu, Jason Wei, Chenyan Jia, and Soroush Vosoughi. 2021a. Modulating language models with emotions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4332–4339, Online. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021b. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware mar-

gIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

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.

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

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*.

Minlie Huang Sahand Sabour, Chujie Zheng. 2021. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. 2021. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3124–3134, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles B Truax and Robert Carkhuff. 2007. *Toward effective counseling and psychotherapy: Training and practice*. Transaction Publishers.

Charles B Truax and Robert R Carkhuff. 1964. Concreteness: A neglected variable in research in psychotherapy. *Journal of Clinical Psychology*.

Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022a. CARE: Causality reasoning for empathetic responses by conditional graph generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 729–741, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022b. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4634–4645, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.