An Iterative Associative Memory Model for Empathetic Response Generation

Anonymous ACL submission

Abstract

Empathetic response generation is to comprehend the cognitive and emotional states in dialogue utterances and generate proper responses. Psychological theories posit that comprehending emotional and cognitive states necessitates iteratively capturing and understanding associated words across dialogue utterances. However, existing approaches regard dialogue utterances as either a long sequence or independent utterances for comprehension, which are prone to overlook the associated words between them.

To address this issue, we propose an Iterative Associative Memory Model (IAMM) for empathetic response generation. Specifically, we employ a novel second-order interaction attention mechanism to iteratively capture vital associated words between dialogue utterances and situations, dialogue history, and a memory module (for storing associated words), thereby accurately and nuancedly comprehending the utterances. We conduct experiments on the Empathetic-Dialogue dataset. Both automatic and human evaluations validate the efficacy of the model. Meanwhile, variant experiments on LLMs also demonstrate that attending to associated words improves empathetic comprehension and expression.

1 Introduction

As an important task for improving dialogue quality, empathetic response generation aims to comprehend the emotional and cognitive states of the user in dialogue utterances and provide appropriate responses (Liang et al., 2021; Liu et al., 2021; Rashkin et al., 2019; Zheng et al., 2021; Zhong et al., 2020).

The majority of methods treat the dialogue utterances as a long sequence to comprehend the user states (Li et al., 2022; Lin et al., 2019; Majumder et al., 2020; Sabour et al., 2022; Zhou et al., 2023). These approaches ignore the discrepancies in meanings among individual utterances, leading to inaccurate understanding of emotional and cognitive states (Wang et al., 2022; Welivita et al., 2021). To address this issue, some methods comprehend more delicate emotional and cognitive states within a set of independent utterances by distinguishing self-other awareness (Zhao et al., 2022) or emphasizing emotion-intent transitions (Wang et al., 2022).

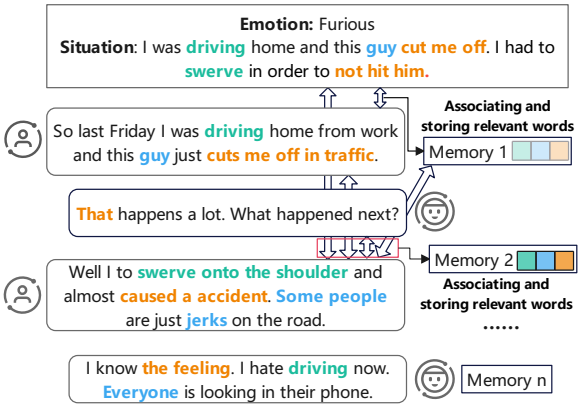

Figure 1: An example of iterative association. Words with the same color are associated. The memory stores the associated words.

However, the situation model (Van Dijk et al., 1983), as an important theory for understanding empathy, posits that comprehending emotional and cognitive states in detail necessitates not only understanding independent utterances, but also iteratively associating pivotal associated words within those utterances (Gernsbacher et al., 1992; Gygax et al., 2003; McNamara and Magliano, 2009; Zwaan and Radvansky, 1998). As shown in Figure 1, the speaker's independent utterances imply the emotion of "anger," yet the overall expression conveys "furious." If the utterances are understood independently, the listener is likely to misinterpret the speaker's emotion as "anger." In contrast, the iterative association integrates subtle associated words, allowing for an accurate understanding of the dialogue. Specifically, when faced with the first utterance, the listener combines this utterance with

related words from the situation and stores it in its memory to form an initial understanding. When encountering the speaker's second utterance, the listener meticulously compares and reasons this utterance with the dialogue history, situation, and related words in memory, to deepen its understanding of the utterance. For instance, associating "jerks" and "guy" reveals an intensification of the emotion of anger, i.e., "furious". Additionally, reasoning that "cut me off" and "caused an accident" makes it easier to realize the speaker's furious due to the life-threatening event. Overall, by associating explicit and implicit information, the listener attains a more nuanced understanding of the utterances. While this comprehension process proves effective, simulating it to achieve meticulous understanding of dialogues remains an open challenge.

In this paper, we propose an iterative associative memory model (IAMM) that iteratively employs an information association module to identify and learn subtle connections within both explicit and implicit information. We first treat the dialogue content, including both the dialogue utterances and situations, as explicit information, and treat the reasoning knowledge about the dialogue content generated by COMET (Hwang et al., 2021) as implicit information. Subsequently, we iteratively utilize the information association module to identify and learn associated words between utterances and situations, dialogue history, and memory (initialized as an empty set) in the explicit/implicit information, and store them in the memory for a thorough understanding of the utterances. Specifically, the information association module, inspired by the idea that "pages (nodes) linked by important pages (nodes) are also more important" (Brin and Page, 1998; Weng et al., 2010), effectively identifies associated words in the to-be-associated sentences through a second-order interaction attention mechanism.

To validate our model, we construct IAMM and IAMM$_{large}$ (LLMs-based model). Experiments are conducted on the Empathetic-Dialogue dataset (Rashkin et al., 2019). Both automatic evaluation and human evaluation demonstrate that compared with the state-of-the-art baselines, our models possess stronger understanding while expressing more informative responses.

Overall, our contributions are as follows:

- We introduce an iterative association framework for empathetic response generation, which simulates the human iterative process of understanding emotions and cognition.
- We propose an iterative associative memory model (IAMM), which iteratively employs a second-order interaction attention mechanism to capture subtle associations in dialogues.
- Experiments on the Empathetic-Dialogue dataset validate the efficacy of our models.

2 Related Work

Empathetic response generation requires comprehending the user states in dialogue utterances to generate appropriate responses (Rashkin et al., 2019). Existing methods can be categorized into dialogue-level models and utterance-level models, according to whether they understand dialogue utterances independently.

Dialogue-level models. Dialogue-level models view all dialogue utterances holistically as a long sequence to comprehend user states. Some dialogue-level models focus on the coarse emotions (Lin et al., 2019; Majumder et al., 2020; Rashkin et al., 2019) or subtle emotions (Gao et al., 2021; Kim et al., 2021; Li et al., 2020, 2022; Yang et al., 2023) present in a conversation to understand the user states. While focusing on emotional states, these models ignore cognitive states, leading to inadequate empathy comprehension (Sabour et al., 2022). Some methods introduce reasoning knowledge to more comprehensively attend to user states (Cai et al., 2023; Sabour et al., 2022; Zhou et al., 2023).

Utterance-level models. Utterance-level models focus on differences in emotional and cognitive states within individual utterances. These models view dialogue utterances as independent sentences to understand the user's state within them via self-other awareness differentiating Zhao et al. (2022) and emotion-intent transitions Wang et al. (2022).

The aforementioned methods view the dialogue utterances as either long sequences or as independent sentences to understand the user's emotions and cognitive states. However, understanding emotional and cognitive states requires iteratively capturing and comprehending associated words (Gernsbacher et al., 1992; Gygax et al., 2003; McNamara and Magliano, 2009). Therefore, we propose an Iterative Associative Memory Model (IAMM), which iteratively understands the associated words between utterances from both explicit

and implicit information, and generates empathetic responses.

3 Method

3.1 Task Formulation

The task of empathetic response generation is as follows: Given the dialogue content, including the context D = [U_1, ..., U_i, ..., U_M] and the situation S = [w_1^s, w_2^s, ..., w_m^s], the model needs to understand the dialogue emotion E and generate an appropriate response Y = [y_1, y_2, ...,y_j, y_N]. U_i = [$w_1^i, w_2^i, ..., w_{m_i}^i$] is the i-th utterance containing m_i words. S represents a situation description consisting of m words. Y represents a response sequence consisting of N words.

3.2 Overview

We illustrate the overall architecture of IAMM in Figure 2. As iterative association requires attention to both explicit and implicit information (Garrod and Sanford, 2012), we view the dialogue content as explicit information, and their reasoning knowledge (Hwang et al., 2021) as implicit information. We construct IAMM in the following three steps: (1) Encoding dialogue information (**Section 3.3**), which learns explicit and implicit information in dialogue through multiple encoders. (2) Capturing associative information (**Section 3.4**), which iteratively captures associated words in the dialogue regarding the explicit and implicit aspects. (3) Predicting emotion and generating responses (**Section 3.5**), which predicts the dialogue emotion and generates responses based on the learned content. Furthermore, we construct a variant model IAMM$_{large}$ (**Section 4.3**), which leverages instructions to guide the large language models to focus on associated words and capture the nuanced associations within the dialogue.

3.3 Encoding Dialogue Information

We encode the dialogue information from both explicit and implicit perspectives.

Explicit Information. We take the dialogue utterances and situations as explicit information. Following previous work (Wang et al., 2022), we add special start tokens [CLS] before the dialogue utterances and situations to obtain word sequences U_i and S, respectively. Based on the word sequence U_i of the dialogue utterances, we first learn sentence representations using an utterance-level encoder Enc_u, then learn the overall meaning of the dialogue using a dialogue-level encoder Enc_{ctx}.

$$H_u^i = Enc_u(E_w^i + E_p + E_r) \quad (1)$$

$$H_c = Enc_{ctx}(H_u^1 \oplus ... \oplus H_u^M) \quad (2)$$

where E_w^i, E_p, E_r are the word embedding, positional embedding, and state embedding of the utterances, respectively. The state embedding is used to distinguish between speakers and listeners. And $H_u^i \in R^{m_i \times d}$, $H_c \in R^{M_D \times d}$, with m_i, H_c being the lengths of the i-th utterance and total dialogue utterances. d and $\oplus$ represent the hidden size and concatenation operation, respectively.

For the word sequence S of the situations, we employ an encoder Enc_{sit} to learn the situation representations.

$$H_s = Enc_{sit}(E_w^{sit} + E_p^{sit}) \quad (3)$$

where E_w^{sit} and E_p^{sit} represent the word embedding and the positional embedding of the situations, respectively. $H_s \in R^{m \times d}$ and m is the number of words it contained.

Implicit Information. Following (Sabour et al., 2022; Wang et al., 2022), we utilize the COMET model (Hwang et al., 2021) to generate reasoning knowledge K_{u_i} and K_s for the dialog utterance U_i and situation S, and take them as the implicit information. Among them, the reasoning knowledge K_{rk} ($rk \in (U_i, s)$) contains 5 types of text: the impact of events (xEffect), personal emotional reactions (xReact), personal intentions (xIntent), personal needs (xNeed), and personal desires (xWant). For convenience, we represent them as K_{rk}^{type}, where $type \in [xEffect, xReact, xIntent, xNeed, xWant]$.

Subsequently, we prepend [CLS] before the reasoning knowledge and feed it into the encoder Enc_{cs}. To understand the overall implicit information, we utilize the encoder Enc_{cs} to learn the reasoning knowledge of the situation and the last utterance.

$$H_{rk}^{type} = Enc_{cs}([CLS] \oplus K_{rk}^{type}) \quad (4)$$

$$H_{rk}^{cs} = \underset{type}{\|} H_{rk}^{type}[0] \quad (5)$$

$$H_{cs} = H_s^{cs} \oplus H_{U_M}^{cs} \quad (6)$$

$$\widetilde{H}_{cs} = \widetilde{Enc}_{cs}(H_{cs}) \quad (7)$$

Where $H_{rk}^{type_i}[0]$ is the overall semantic representation of the sentence, and $rk \in (U_i, s)$, $type \in [xEffect, xReact, xIntent, xNeed, xWant]$. $\|$

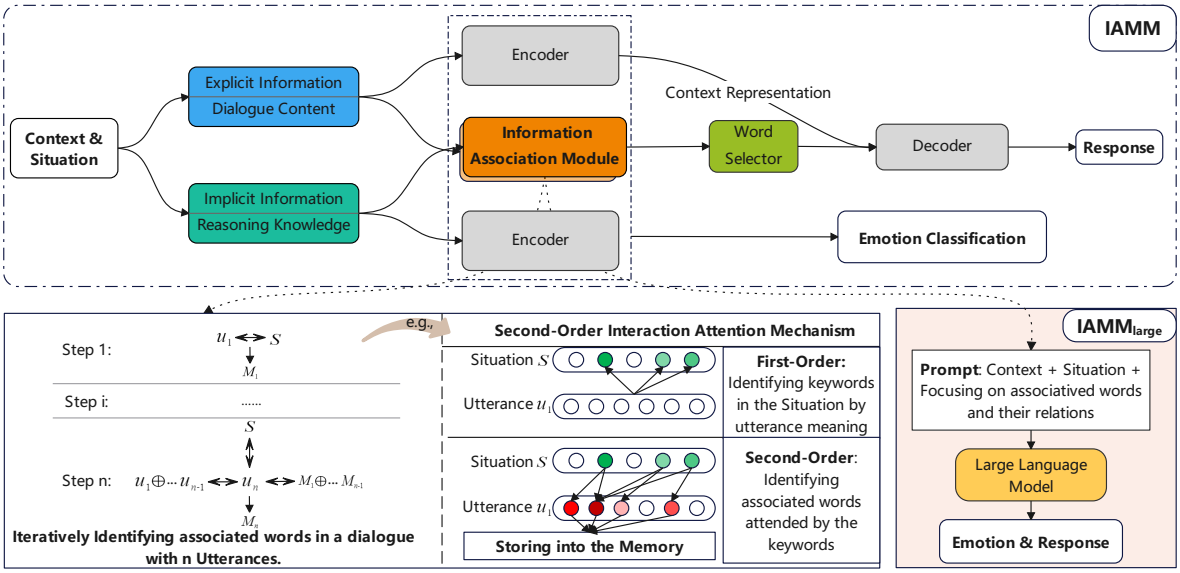

Figure 2: Overview of IAMM and IAMM$_{large}$, which are small-scale and large-scale models focusing on association information, respectively. IAMM mainly consists of the following steps: (1) Encoding dialogue information, including explicit information (dialogue utterances and situation) and implicit information (reasoning knowledge); (2) Iteratively capturing associative information, namely associated words between sentences; (3) Predicting emotion and generating responses. Moreover, IAMM$_{large}$ focuses on subtle associations by injecting associated words into instructions.

is the concatenation operation. U_M and s represent the last utterance and situation, respectively. $H_{rk}^{type} \in R^{d_{rk}^{type} \times d}$, $H_{cs} \in R^{10 \times d}$, $\widetilde{H}_{cs} \in R^{10 \times d}$ and d_{rk}^{type} is the number of words in the reasoning knowledge K_{rk}^{type}.

3.4 Capturing Associated Information

We employ the information association module (IAM) to capture important associated words in explicit and implicit information to coherently and thoroughly understand the dialogue. To elaborate on the process of the information association module, we explain it in two aspects: Iterative association process, which describes the process executed by the information association module. Information association module, which explicates in detail the internal structure of the module.

3.4.1 Iterative Association Process

To understand the current utterance U_i, we construct pairs of explicit and implicit information as the input to the information association module. These pairs are used to capture important associations between utterance-situation, utterance-dialogue history, and utterance-memory, respectively. Here, M_i and $\overline{M}_i$ represent the memory, which is used to store associated words and is initialized as empty. Note that when U_i=U_1, both the dialogue history and memory are empty.

Explicit Information Pairs: (E_w^i, E_s), $(E_w^i, E_w^1 \oplus ..., E_w^{i-1})$, $(E_w^i, M_1 \oplus ..., M_{i-1})$.

Implicit information pairs: $(H_{U_i}^{cs}, H_s^{cs})$, $(H_{U_i}^{cs}, H_{U_1}^{cs} \oplus ..., H_{U_{i-1}}^{cs})$, $(H_{U_i}^{cs}, \overline{M}_1 \oplus ..., \overline{M}_{i-1})$.

Next, we feed the pairs of explicit and implicit information into the information association module (IAM). Since the processing procedures for the two types of information are the same, we take the processing procedure for explicit information as an example to illustrate.

Based on the utterance pairs of U_i, including utterance-situation, utterance-dialogue history, and utterance-memory, the IAM bidirectionally identifies key associated words in the pairs' sentences and stores them in the memory. Subsequently, the model continuously capture the associated words of the next utterance U_{i+1} until the last dialogue utterance. This iterative process facilitates thoroughly capturing the intricate associations between the utterance and related sentences, thereby enabling better capture of critical information.

3.4.2 Information Association Module

This module identifies associated words between relevant sentences, such as the situation and utterance. Inspired by the notion that "nodes highly attended by key nodes are likely important", we pro-

pose a **second-order interaction attention mechanism** with two processes: the first-order interaction attention identifies situation words attended by the overall utterance meaning as keywords (key nodes). The second-order interaction attention identifies utterance words attended by the keywords as associated words (the nodes highly attended by key nodes). This two-stage filtering selects associated words that explicitly reflect the connections between the two sentences.

Since associated words exist simultaneously in two sentences (e.g. in the utterance or situation), we bidirectionally select words from the sentences. For example, selecting associated words in the situation based on the utterance (i.e., u2s), and selecting associated words in the utterance based on the situation (i.e., s2u). For simplicity, we elaborate on the whole process using the example of s2u: (1) Constructing association matrices that contain association scores between words and are used for word identification. (2) First-order Interaction Attention, which identifies keywords in the situation. (3) Second-order Interaction Attention, which identifies associated words in the utterance based on the keywords in the situation. (4) Storing the bidirectionally associated words in memory for the next processing.

Constructing Association Matrices. To capture rich features, we construct two multi-head association matrices. A^{s2t} represents the association scores from the situation words to the utterance words, while A^{t2s} represents the association scores from the utterance words to the situation words.

$$A_{ij}^{s2t} = \mathop{\Big\|}_{n=1}^{H} \sigma((w_q^n v_{s,i}^n)^T (w_k^n v_{t,j}^n)) \tag{8}$$

$$A_{ij}^{t2s} = \mathop{\Big\|}_{n=1}^{H} \sigma((w_q^n v_{t,i}^n)^T (w_k^n v_{s,j}^n)) \tag{9}$$

where $A^{s2t} \in R^{H \times d_s \times d_t}$, $A^{t2s} \in R^{H \times d_t \times d_s}$ and H is the number of multi-heads. Here, σ is the Sigmoid function.

First-order Interaction Attention. We identify keywords in the situation. Intuitively, the words that are associated as important by most other words are more likely to be important. Therefore, we select the words in the situation that have higher degrees of association with utterance words as the keywords.

$$A^s = Mean(A^{t2s}) \tag{10}$$

$$A_{k_1}^s = \widetilde{Top}_{k_1}(A^s) \tag{11}$$

where $Mean$ and $\widetilde{Top}$ are the average function and filter function respectively. k_1 represents the number of important words to filter. And $A^s \in R^{H \times d_s}$, $A_{k_1}^s \in R^{H \times k_1}$.

Second-order Interaction Attention. We select words in the utterance that have the highest degrees of association with the key situation words, as important associated words.

$$A_{k_2}^{s2t}, E_{k_2}^{s2t} = Top_{k_2}(A^{t2s}) \tag{12}$$

$$\widetilde{E}_{k_2}^{s2t} = \mathop{\Big\|}_{k=1}^{k_2} A_k^{s2t} \cdot E_k^{s2t} \tag{13}$$

$$E_t^{st} = A_{k_1}^s \cdot \widetilde{E}_{k_1,k_2}^{s2t} \tag{14}$$

where $A_{k_2}^{s2t} \in R^{H \times d_s \times k_2}$, $E_{k_2}^{s2t} \in R^{H \times d_s \times k_2 \times d_h}$, and they are the representations and scores of the utterance words attended by situation words, respectively. $K2$ is the number of associated words. $\widetilde{E}_{k_2}^{s2t} \in R^{H \times d_s \times (k_2 \times d_h)}$, $E_t^{st} \in R^{H \times k_1 \times (k_2 \times d_h)}$ and d_h is the hidden size. E_t^{st} is the representation of associated words.

We select the associated words from the situation in the same way. Subsequently, we concatenate the associated words of the situation and the utterance, where $E_{st} \in R^{(2H \times k_1) \times (k_2 \times d_h)}$.

$$E_{st} = E_t^{st} \oplus E_s^{st} \tag{15}$$

Storing the Associated Words in Memory. Regarding explicit information, we select associated words between utterances and situations (E_{sc}^{ek}), between utterances and dialogue history (E_{cc}^{ek}), and between utterances and memory (E_{mc}^{ek}) in the same way.

$$V_{ek} = E_{sc}^{ek} \oplus E_{cc}^{ek} \oplus E_{mc}^{ek} \tag{16}$$

where $E_{sc}^{ek}, E_{cc}^{ek}, E_{mc}^{ek} \in R^{(2H \times k_1) \times (k_2 \times d_h)}$.

Specifically, regarding implicit information, we also construct associated words between utterances and situations (E_{sc}^{ik}), between utterances and dialogue history (E_{cc}^{ik}), and between utterances and memory (E_{mc}^{ik}). When iterating to the last sentence, we combine the explicit information memory V_{ek} and the implicit information memory V_{ik} as the final associated information V.

$$V_{ik} = E_{sc}^{ik} \oplus E_{cc}^{ik} \oplus E_{mc}^{ik} \tag{17}$$

$$V = V_{ek} \oplus V_{ik} \tag{18}$$

where $E_{sc}^{ik}, E_{cc}^{ik}, E_{mc}^{ik} \in R^{(2H \times k_1) \times (k_2 \times d_h)}$. $V \in R^{L_m \times d}$ and L_m is the number of associated words in the memory.

Based on the memory of associated words, we learn association information through an encoder Enc_m.

$$H_m = Enc_m(V) \tag{19}$$

3.5 Predicting Emotion and Generating Response

3.5.1 Prediction Emotion

In order to predict emotion, we input the dialogue utterance representation H_c, the situational representation H_s, into aggregation network AN_u (Yang et al., 2023) to obtain emotion representations. We subsequently use these representations to separately predict the probability of emotions.

$$P_c = \phi(AN_u(H_c)) \tag{20}$$

$$P_s = \phi(AN_u(H_s)) \tag{21}$$

where ϕ is the softmax function. $P_c, P_s \in R^{d_e}$ and d_e is the number of emotions.

Similarly, we also predict the emotion probabilities for the association representation H_m, and the reasoning knowledge representation H_{cs}.

$$P_a = \phi(AN_a(H_m)) \tag{22}$$

$$P_{cs} = \phi(AN_{cs}(\widetilde{H}_{cs})) \tag{23}$$

where $P_a, P_{cs} \in R^{d_e}$. AN_a, and AN_{cs} represent aggregate networks with the same architecture but different parameters.

We multiply the above emotion probabilities as the final emotion probability. Then we use log-likelihood loss to optimize the parameters based on the emotion probability and the ground truth label e^*.

$$P_e = P_c(e^*) \cdot P_s(e^*) \cdot P_a(e^*) \cdot P_{cs}(e^*) \tag{24}$$

$$\mathcal{L}_e = -log(P_e) \tag{25}$$

3.5.2 Generation Response

To fully utilize the associative words containing important information, we design a word selector and flexibly incorporate associative information during decoding.

Word Selector. We select important associated words to utilize the effective information in memory.

$$S_m = \sigma(w_v H_m) \tag{26}$$

$$\widetilde{S}_m, \widetilde{H}_m = Top_{k_3}(S_m, H_m) \tag{27}$$

$$\widetilde{H}_m = \widetilde{S}_m \cdot \widetilde{H}_m \tag{28}$$

where Top is a selecting function that selects the top associated words with the highest scores from the memory H_m based on the score S_s. $w_v \in R^{d\times 1}$ is a trainable parameter. $\widetilde{H}_m \in R^{k_3 \times d}$ and K_3 is a hyperparameter.

Based on the representations of the utterance and associated words, we generate the decoding vectors O_c and O_s , respectively.

$$O_c = Dec_c(H_c) \tag{29}$$

$$O_m = Dec_a(\widetilde{H}_m) \tag{30}$$

where Dec_c and Dec_a represent decoders with the same architecture but different parameters. O_c, $O_m \in R^{L_t \times d}$, and L_t is the length of the response words at time t.

We then combine the decoding vectors to form O and use it to predict word probabilities.

$$g = \sigma(w(O_c \oplus O_m)) \tag{31}$$

$$O = g \cdot O_c + (1 - g) \cdot O_m \tag{32}$$

$$P(y_t|y < t, D, S) = Generator(E_{y<t}, O) \tag{33}$$

where $w \in R^{d\times 1}$ represents learnable parameters. $Generator$ denotes the point generator (See et al., 2017) that transforms the decoded vectors into word probabilities.

Finally, we employ cross-entropy loss as the generation loss $\mathcal{L}_{gen}(y_t)$. By integrating the generation loss $\mathcal{L}_{gen}(y_t)$ and the emotion loss $\mathcal{L}_e$, we optimize the overall parameters.

$$\mathcal{L}_{gen}(y_t) = -\sum_{t=1}^{T} log(P(y_t|y < t, D, S)) \tag{34}$$

$$\mathcal{L} = \mathcal{L}_{gen}(y_t) + \mathcal{L}_e \tag{35}$$

3.6 Baselines

To compare the performance of IAMM, we select the state-of-the-art models as baselines. **EmpDG** (Li et al., 2020) considers fine-grained emotional words and user feedback; **KEMP** (Li et al., 2022) enhances hidden emotional representations using a ConceptNet-based emotional graph; **CEM** (Sabour et al., 2022) enhances emotion and cognition using reasoning knowledge; **CASE** (Zhou et al., 2023) aligns emotion and cognition from fine-grained and coarse-grained perspectives; **SEEK** (Wang et al., 2022) is a model that captures emotional-intention transitions in dialogue utterances; **ESCM** (Yang et al., 2023) captures emotion-semantic dynamic associations based on word-level emotions.

Models	Acc	PPL	Dist-1	Dist-2
EmpDG	34.31	37.29	0.46	2.02
KEMP	39.31	36.89	0.55	2.29
CEM	39.11	36.11	0.66	2.99
CASE	40.2	35.37	0.74	4.01
SEEK	41.85	37.09	0.73	3.23
ESCM	41.19	**34.82**	1.19	4.11
IAMM	**55.92**	35.66	**2.09**	**7.03**
w/o EA	52.43	35.17	1.18	4.04
w/o IA	51.48	35.29	1.54	5.14
w/o WS	55.82	35.46	1.75	6.01

Table 1: Results of automatic evaluation.

Comparisons	Aspects	Win	Lose	κ
IAMM vs. ESCM	Emp.	**23.5**	20.0	0.43
	Rel.	**24.4**	22.7	0.42
	Flu.	**18.0**	16.8	0.43
IAMM vs. SEEk	Emp.	**33.6**	20.2	0.46
	Rel.	**41.2**	16.8	0.45
	Flu.	**18.5**	16.5	0.43

Table 2: Results of human evaluation, where κ is the inter-labeler agreement measured by Fleiss's kappa (Fleiss and Cohen, 1973), and $0.4 < \kappa \leq 0.6$ indicates moderate agreement.

3.7 Implementation Details

We conduct experiments on the EMPATHETIC-DIALOGUES (Rashkin et al., 2019) dataset. The details are provided in Appendix A.

3.8 Evaluation Metrics

To comprehensively understand model performance, we conduct automatic and human evaluations.

Automatic Evaluation. Following previous methods (Li et al., 2022; Sabour et al., 2022), we use perplexity (PPL), accuracy (Acc), distinct-1 (Dist-1)/distinct-2 (Dist-2) (Li et al., 2015) to evaluate response fluency, emotion classification accuracy, and response diversity, respectively. Specifically, lower perplexity indicates better quality, while higher values are better for the other metrics.

Human Evaluation Metrics. We invite three professional crowdworkers to evaluate the response quality. Consistent with previous methods (Majumder et al., 2020; Yang et al., 2023), we conduct A/B testing to compare the baseline and IAMM. For a response, if IAMM has better quality, the crowdworkers add one point to *Win*. If IAMM is worse than the comparison model, they add one point to *Lose*. To evaluate quality, we consider empathy (Emp.), relevance (Rel.), and fluency (Flu.): Empathy measures whether the emotion in the response is appropriate. Relevance measures whether the response is relevant to the dialogue topic and content. Fluency measures whether the language of the response is natural and fluent.

4 Results and Analysis

4.1 Main Results

Automatic Evaluation Results. As shown in Table 1, IAMM outperforms the baselines on most metrics. For emotion accuracy, IAMM significantly surpasses the baselines. This is because leveraging associated information facilitates dialogue subtle understanding, which promotes accurate emotional comprehension. In diversity, IAMM substantially surpasses the baselines. This is primarily attributed to the focused important associated words facilitating the generation of informative responses. Regarding perplexity, IAMM does not surpass the baseline. This is because the responses generated by the model contain fewer general sentences, which have better perplexity and lower diversity.

Human Evaluation Results. As shown in Table 2, IAMM also demonstrates better performance over the baselines in human evaluation. The model shows superiority in empathy. This is because focusing on the associations between utterances promotes delicate understanding, thus expressing appropriate emotional responses. The model is excellent in relevance. This is primarily because associated words tend to be important words, and responses generated based on these significant words are more likely to be relevant. Regarding fluency, IAMM is better than the baselines in human evaluation, although it is worse on the automatic metric (PPL). This is primarily because the automatic metric (PPL) tends to favor sentences that are common in the training set, while human evaluation conforms better to natural language. IAMM generates fewer common and general sentences, such as "I am so sorry to hear that." Hence, the evaluation results turn out this way.

4.2 Ablation Studies

As shown in Table 1, we conduct the following ablation experiments to validate the effectiveness of each module: **w/o EA**: without explicit association; **w/o IA**: without implicit association; **w/o WSD**:

Models	Acc	Dist-1	Dist-2
$GLM3_{7B}$	62.16	3.51	21.56
$IAMM_{GLM}$	**62.6**	**3.55**	**21.7**
GPT3.5	37.9	3.58	21.38
$IAMM_{GPT}$	**38.51**	**3.63**	**22.13**

Table 3: Automatic evaluation results of IAMM on large language models.

without the associated word selector and decoder.

For emotion accuracy and diversity, the results show both explicit and implicit associative information have considerable influence. Explicit associative information contributes more to diversity, while implicit information contributes more to emotion inference. This indicates that associated words in explicit information are more easily utilized for expression, while those in implicit information are more conducive for emotion inference. For perplexity, incorporating both types of information leads to better perplexity, as more non-general informative expressions are generated. Furthermore, without the associated word selector and decoder, the model's diversity decreases, owing to insufficient expression of key information.

4.3 IAMM on Large Language Models

Large language models have shown superior performance on multiple tasks (Chen et al., 2023; Qin et al., 2023; Sun et al., 2023; Tang et al., 2023; Wang et al., 2023; Zheng et al., 2023). To further verify the effect of iterative association on large models, we first make the large model pay attention to the associated words captured by IAMM. We then use instructions to focus on the relationships between the associated words to deeply understand the dialogue. The verification methods include fine-tuning and non-fine-tuning. Regarding the fine-tuning method, we input the associated words through instructions on the Chinese-English mixed model ChatGLM3 (Du et al., 2022; Zeng et al., 2022) and conducted fine-tuning training. Regarding the non-fine-tuning method, we instruct GPT-3.5 to pay attention to the associated words and their relationships to enhance the model.

The experimental results are shown in Table 3. Compared with the baseline models, the models focusing on associative relationships have stronger emotion recognition and expression abilities, which further demonstrates the effectiveness of iterative associations.

4.4 Analysis of Associated Words

To further explore the characteristics of associated words, we collected 8,012 associated words on the test set and statistically analyzed their emotion intensity and inverse document frequency (IDF) [1]. The analysis results show: (1) The most frequently attended words are common words (e.g., "that", "guys"). (2) Words that are paid more attention to (assigned higher weights) have either greater emotional intensity or are non-stop words (e.g. "traffic", "accident"). The main reasons are: (1) The model tends to correlate phrase-phrase and phrase-word associations, such as (swerve onto the shoulder, caused a accident) and (cuts me off in traffic, That) in Example 1. In the phrase-word associations, phrases often contain both common and uncommon words, while the individual words are typically common words. This results in more common words attended by the model. (2) The model places more emphasis on phrases with emotion or definite meanings. To better understand the emotion or meaning, the model often assigns higher weights to emotion words or uncommon words within the phrase. See Appendix C for details.

4.5 Case Study

We conduct case studies for the strongest baseline and IAMM. See Appendix B for details.

5 Conclusion and Future Work

In this paper, we have proposed an Iterative Associative Memory Model (IAMM) for empathetic response generation, inspired by the human iterative process of understanding emotions and cognition. It employs a novel second-order interaction attention mechanism to iteratively identify key associated words across dialogue utterances, enabling a more accurate understanding of the emotional and cognitive states. Automatic and human evaluations demonstrate that IAMM accurately understands emotions and expresses more empathetic responses. Experiments based on large language models and associated word analysis further validate the effectiveness of the iterative associations. In the future, we will explore empathetic comprehension mechanisms based on large language models.

[1] Characteristics of IDF: The more common a word is, the lower its IDF score, and vice versa.

6 Limitations

The limitations of our work are as follows: (1) This work is inspired by the text-based empathetic comprehension mechanism. As the better comprehension of empathy relies on multimodal and large language models, we will conduct research combining these aspects in the future. (2) Iterative association relies on situation information. Although it is prevalent and effective, some datasets still lack this feature. In the future, we will also explore how to effectively construct situation information.

7 Ethical Considerations

Regarding the potential ethical impacts of our work: (1) The dataset we use is EMPATHETIC-DIALOGUE, which is open source and does not involve any potential ethical risks. (2) The baseline models we use are also public and do not have potential moral impacts. Moreover, the components employed in our model are open-sourced or innovative and do not involve potential ethical risks.

References

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. *arXiv preprint arXiv:2306.04657*.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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 EMNLP*, pages 807–819.

Simon Garrod and Anthony Sanford. 2012. Referential processes in reading: Focusing on roles and individuals. In *Comprehension processes in reading*, pages 465–486. Routledge.

Morton Ann Gernsbacher, H Hill Goldsmith, and Rachel RW Robertson. 1992. Do readers mentally represent characters' emotional states? *Cognition & Emotion*, 6(2):89–111.

Pascal Gygax, Jane Oakhill, and Alan Garnham. 2003. The representation of characters' emotional responses: Do readers infer specific emotions? *Cognition and Emotion*, 17(3):413–428.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, volume 35, pages 6384–6392.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *EMNLP*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:abs/1412.6980*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv:abs/1510.03055*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv:abs/1911.08698*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *AAAI*.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In *AAAI*, volume 35, pages 13343–13352.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, page 121–132.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv:abs/2106.01144*.

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 *EMNLP*, page 8968–8979.

Danielle S McNamara and Joe Magliano. 2009. Toward a comprehensive model of comprehension. *Psychology of learning and motivation*, 51:297–384.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, page 5370–5381.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Virginia, USA. AAAI Press.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Teun Adrianus Van Dijk, Walter Kintsch, et al. 1983. Strategies of discourse comprehension.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715*.

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.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Wu Yunbing, Yisong Su, Sibo Ju, and Xiangwen Liao. 2023. Exploiting emotion-semantic correlations for empathetic response generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2022. Don't lose yourself! empathetic response generation via explicit self-other awareness. *arXiv preprint arXiv:2210.03884*.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *ACL*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv:abs/2004.12316*.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237, Toronto, Canada. Association for Computational Linguistics.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162.

A Appendix A

We conduct experiments on the EMPATHETIC-DIALOGUES (Rashkin et al., 2019) dataset, which contains 32 types of emotions, i.e. d_e = 32. In the encoder, the hidden size is d = 300. In the information association module, the numbers of heads and hidden layer size are H = 2 and d_h = 20, respectively. The numbers of keywords and associated words are k_2 = 15 and k_1 = 5, respectively. And the number of selected keywords for the decoder is k_3 = 5. The batch size of the model is set to 16. We optimize the model using Adam optimizer (Kingma and Ba, 2014) on an NVIDIA Tesla T4 GPU. The model converges after 14,400 iterations.

B Appendix B

We select the strongest baselines to compare with IAMM. The details of the results are listed in Table 4.

In the first case, SEEK fails to correctly express the emotion of “proud”. ESCM incorrectly recognizes the emotion within. While IAMM perceives the speaker’s feeling of “proud” by identifying the associative words “accepted into harvar”. At the same time, it also understands the subject stated by the speaker, “my daughter”, thus incorporates emotions and responses appropriately in the response “bet she was so proud of her”.

In the second case, as the models fail to identify key information in the dialogue, SEEK and ESCM express general sentences. While IAMM discovers the key information “family” in the dialogue by associating “my family” and “they”. At the same time, by associating the emotion "ashamed" in the sentence, IAMM also clearly understands the emotion. Based on the key information and emotion, the model generates an empathetic response.

Overall, the iterative associated words in dialogues facilitate nuanced understanding. Additionally, by conveying key associated words, IAMM produces more informative and relevant responses.

C Appendix C

To validate the characteristics of the associated words, we extracted 8,012 associated words from the test set that the model paid attention to. We statistically analyzed the data in terms of emotions and word frequency.

Figure 3 shows the results of the emotion analysis. The red line ranks words by attention counts from high to low, and calculates the average emotion intensity of the top k words. The blue line ranks words by the weight given by the model, and calculates the emotion intensity of the top k words. The gray line is the average emotion intensity of all words in the test set. The x-axis is the number of top k words, and the y-axis is the emotion intensity value. The red line indicates that the most attended words by the model have relatively low emotion intensity, while the blue line indicates that the most attended words by the model have high emotion intensity.

Figure 4 shows the results of word frequency analysis. The red line sorts words by attention counts and calculates the average inverse document frequency (IDF) of the top k words. The blue line sorts words by the weight given by the model, and the gray line is the average IDF for all words. Similarly, the x-axis represents the number of top k words, and the y-axis represents the IDF value. The red line indicates that the most attended words are the most common ones, while the blue line indicates that the most attended words are uncommon.

The overall results show that the model pays attention to common words with low emotions (e.g. "That", "it"), while its most highly weighted words have high emotion intensity or are less common.

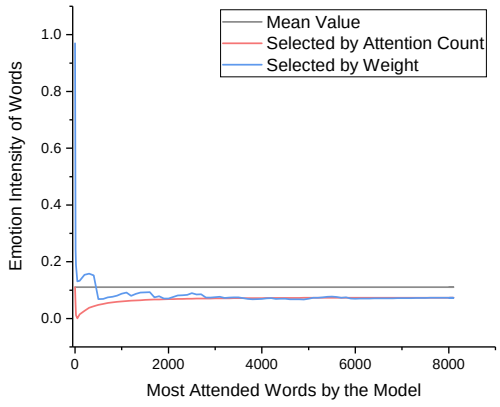

Figure 3: Results of emotion analysis for associated words.

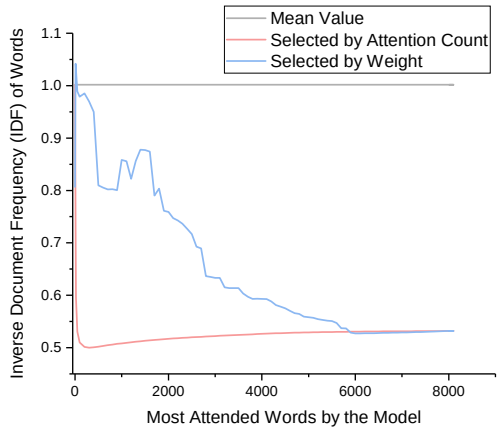

Figure 4: Results of frequency analysis for associated words.

Emotion	**Proud**
Situation	I was so excited of my daughter when she was accepted into harvard. She worked so hard through high school and has achieved her dream.
Speaker	My daughter was accepted into harvard. She starts in the fall.
Gold	Any employer would hire someone who went to harvard in a heartbeat. You are a lucky parent!
SEEK	Wow, that is a very long time.
ESCM	Oh wow, that is a shame.
IAMM	That is awesome! I bet she was so proud of her!
Emotion	**Ashamed**
Situation	I am very ashamed in my family.
Speaker	I am very ashamed in my family.
Listener	I am really sorry to hear that. Why are you ashamed?
Speaker	Because they never invite me to go places, and they never call me.
Gold	That is really too bad. Did you all have a fight?
SEEK	I am sure you will do well.
ESCM	I am sorry to hear that.
IAMM	I know that feeling. I have a family and I know how you feel.

Table 4: Case Study of IAMM and Benchmarks, where words with the same color have certain association relationships.