The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

基於常識知識的移情對話回覆生成
Improving Response Diversity through Commonsense-Aware Empathetic Response Generation

黃紫嫻 **Tzu-Hsien Huang**
中央大學資訊工程學系
christy514@g.ncu.edu.tw

張嘉惠 **Chia-Hui Chang**
中央大學資訊工程學系
chia@csie.ncu.edu.tw

摘要

本篇論文著重在移情對話生成任務上。先前研究關於移情對話生成的方法 (Majumder et al., 2020b; Lin et al., 2019) 主要集中在檢測和利用用户的情緒來產生移情反應。本研究將使用額外的常識知識圖譜做爲機器人對常識性的背景知識。我們針對非預訓練和預訓練模型各使用不同的方式增強多樣性，在非預訓練模型上我們將 AdaLabel(Wang et al., 2021) 應用在 CEM 模型 (Sabour et al., 2022) 上，而對於預訓練模型我們使用 BART 模型結合多種常識知識讓模型能生成更有資訊的移情回應。研究結果顯示所提出的模型在 EMPATHETICDIALOGUES 和 DailyDialog 資料集上都優於基線模型，並且在個案研究中可以看到模型產生更多信息和同理心的回應。

Abstract

Due to the lack of conversation practice, the main challenge for the second-language learners is speaking. Our goal is to develop a chatbot to encourage individuals to reflect, describe, analyse and communicate what they read as well as improve students' English expression skills. In this paper, we exploit COMMET, an inferential commonsense knowledge generator, as the background knowledge to improve the generation diversity. We consider two approaches to increase the diversity of empathetic response generation. For non-pretrained models, We apply AdaLabel (Wang et al., 2021) to Commonsense-aware Empathetic model (Sabour et al., 2022) and improve Distinct-2 score from 2.99 to 4.08 on EMPATHETIC DIALOGUES (ED). Furthermore, we augment the pretrained BART model with various commonsense knowledge to generate more informative empathetic responses. Not only has the automatic evaluation of distinct-2 scores improved from 9.11 to 11.21, but the manual case study also shows that CE-BART significantly outperform CEM-AdaLabel.

關鍵字：移情對話回應生成、知識感知回應生成、回應多樣性

Keywords: Empathetic response generation, Commonsense aware response generation, response diversity

1 緒論

提升學生英語口說能力的方法之一是透過大量的對話練習，因此建構一個聊故事機器人，讓同學們透過有主題的故事和機器人進行聊天，是教育型對話機器人的一個重要研究方向 (Liu et al., 2022)。我們的目標是開發一個聊故事機器人，在英文閱讀活動中讓機器人扮演陪伴與支持的角色，藉由和學生一同談論英文故事書、分享心得，促進其英文閱讀之興趣發展。

聊故事機器人雖然可以由機器人主導對話的進行，但是當學習者回答之後，如何接續使用者的回應是一個相當大的挑戰。爲了產生合理的回應，近年的對話系統研究試圖應用知識圖譜來豐富對話回應內容，例如 KEMP(Li et al., 2022) 及 MIME(Majumder et al., 2020b) 引用 ConceptNet 於 Empathetic Dialogues (ED) 資料集上的對話回應生成，彌補機器人背景知識的不足，讓對話系統可以生成包含同理心的回應，合理地回應他人的情況和感受，避免對話過於單調和死板。

雖然 ConceptNet 包含多語言的詞彙知識、以及不同詞彙之間的關係 (例如：UsedFor, RelatedTo, Antonym)、也有物理常識知識 (例如：HasA, PartOf)，然而 ConceptNet 對於事件描述涵蓋有限，僅 800 萬的節點中不到百分之一的節點屬於事件，而 2,100 萬鏈結中僅有不及千分之六的鏈結是關於事件。因此著重於在"If-Then" 推論知識的 ATOMIC(Sap et al., 2019) 的推出送到相當大的關注。同時通過微調預訓練語言模型 (GPT-2(Radford

et al., 2018)、BART(Lewis et al., 2020))，Bosselut 等人更提出 COMET(Bosselut et al., 2019) 常識推論模型，可以生成更多不存在於原始知識庫的常識知識。

Sabour et al.即應用 COMET 生成的常識，有效增加 ED 對話集回應的新穎單字 (Dist-1) 及雙字 (Dist-2) 多樣性到 0.66% 及 2.99%。本篇論文著重在結合常識知識於移情對話生成，並提高模型生成多樣性，針對非預訓練模型和預訓練模型兩種模型使用不同增強生成多樣性的方法。本篇論文貢獻有以下幾點：

- 我們成功將一個可以自適應估計目標標籤分布的方法 AdaLabel 應用在 CEM 模型上面，自動評估結果表明，與 CEM 模型的方法相比，CEM-AdaLabel 可以分別提升 Dist-1 及 Dist-2 至 0.79% 及 4.08%。
- 我們提出 CE-BART 透過預訓練的 BART 模型和 5 種類型的常識來增強同理心反應生成的方法，同時提升 Dist-1 及 Dist-2 至 2.35% 及 11.21%。
- CE-BART 在自動評估和人工評估上都取得很高的效能，並且在個案研究（Case Study）結果表明 CE-BART 可以產生更多信息和同理心的反應。

2 相關研究

2.1 常識知識圖與知識擷取生成

現有常識知識包括 WordNet, ConceptNet, FrameNet, Atomic, 等常識知識庫。ConceptNet(Liu and Singh, 2004) 是一個多語言的知識庫，主要表示單詞或短語之間的常識關係。此知識庫是以起始節點、關係標籤和結束節點的三元組形式表示一段關係 (例: A *net* is used for *catching fish*，可以表示成 (*net*, UsedFor, *catching fish*))。在 ConceptNet (v5.7)(Speer et al., 2017) 中資料來源爲透過眾包收集並與 Wikitionary、WordNet、OpenCyc 和 DBPedia 等現有知識庫合併，總共包含 2100 萬個邊、800 萬個以上的節點和 36 個關係 (relation)，主要關注在詞彙知識 (例如：UsedFor, RelatedTo, Antonym) 和物理常識知識 (例如：HasA, PartOf)，所以 ConceptNet 偏向詞與詞之間的關係，對於事件描述涵蓋有限。

ATOMIC (An atlas of machine commonsense)(Sap et al., 2019) 則是著重在"If-Then"的推論知識，收集了約 88 萬個以上的推理知識實例。"If-Then" 關係可以分成三大類：(1) 事件導致心理狀態 (If-Event-Then-Mental-State)，(2) 事件導致事件 (If-Event-Then-Event)，(3) 事件導致表像人格 (If-Event-Then-Persona)，共提供九種關係類別: oEffect、oReact、oWant、xAttr、xIntent、xNeed、xEffect、xReact 和 xWant，"x" 代表事件和原因是發生在人身上，而"o"則是發生在其他人身上。不同於其他知識庫，ATOMIC 中的節點形式爲 free-text 的方式，所以可以在日常常識方面更具表現力。

因開放式對話中所包含的常識是無限的，可能會有事件無法對應到現有知識庫的狀況，所以 Bosselut 等人 (Bosselut et al., 2019) 提出 COMET 模型，通過在現有常識知識庫上微調預訓練語言模型 (GPT-2(Radford et al., 2018)、BART(Lewis et al., 2020))，此模型可以生成更多不存在於原始知識庫的常識知識。

COMET 模型可以爲任何事件或短語按照需要的關係生成常識知識，這種靈活性使他可以快速的應用在許多任務中，近年常被用於對話回覆任務像是基於角色的對話 (Majumder et al., 2020a) 和移情對話 (Ghosal et al., 2020; Sabour et al., 2022; Zhu et al., 2021)。

2.2 移情對話回應生成

同理心是人類日常對話中重要的技巧，它使人可以感知、理解和適當地回應他人的情況和感受。早期移情對話系統的研究 (Wang and Wan, 2018; Zhou et al., 2018) 比較集中在特定情緒下產生反應，但在近期研究中認爲移情更重要的是去理解對話中說話者的情感並產生包含同理心的回覆，偵測說話者的情感對於生成移情回應是必須的 (Rashkin et al., 2019)。Lin 等人 (Lin et al., 2019) 提出 MoEL 模型爲每種情緒設計單獨的解碼器，並"*softly combine*"解碼器的輸出，這樣的設計可以明確地學習如何根據對上下文情感的理解來選擇適當的反應; Majumder 等人 (Majumder et al., 2020b) 認爲移情反應需要模仿說話者的情緒，回應的情緒除了會與說話者一致外，有時也會是包含正負面的情緒，所以作者將情緒分成消極和積極兩組，在訓練過程中適當地組合來平衡用户情感的模仿，相較於 MoEL，此模型因包含多樣情緒所以可以生成更多樣的回覆。

爲了讓對話系統的回覆能更具同理心和人性化，近期研究更是將理性與情感結合，使得生成的回覆除了包含信息外還能有適當的情緒，讓用户得到更滿意的回覆。Zhong 等人 (Li et al., 2021) 提出 CARE 模型並建構了一個基於情感的常識知識圖譜 (EA-CKG)，使用與 ConceptNet 的 n-gram 匹配來從 message 和 response 中提取 concept，情感三元組被定義爲 {message concept, emotion, response concept}，除了用匹配提取的方式，此論文在 EA-CKG 上訓練一個知識

Dialogue History

Speaker: I feel like a terrible sibling right now .
Listener: What did you do to feel that way ?
Speaker: My sister , who lives out of state and I do not see often , was recently in town visiting our dad . I did not visit with them .

Response

I am sorry to hear that , you could make it up to them by going to visit ?

Figure 1: Empathetic Dilogues (ED) 資料集對話範例。

嵌入模型 TransE(Bordes et al., 2013) 來學習全局概念和關係的嵌入，並透過關聯性的計算來擷取任意數量的新的相關概念，最後將 EA-CKG 構建的常識和情感合併到基於 Transformer(Vaswani et al., 2017) 的回覆生成模型中，與 ConceptFlow(Zhang et al., 2020) 不同的是，CARE 不受常識庫的範圍限制，它可以產生新的常識來生成回覆。Li 等人 (Li et al., 2022) 也在同年提出 KEMP 模型，利用歷史對話、ConceptNet 和情感辭典 NRC_VAD(Mohammad, 2018) 來構建情緒知識圖，相較於 CARE 是預先將所有對話構建成一個基於情感的常識知識圖譜，KEMP 是針對每個對話提取對應到 ConceptNet 中更高情感強度值的 concept 來構成圖譜，並使用 multi-head graph attention 來更新 node，故可以增強移情對話生成模型的情感感知。

3 回應生成任務描述

我們首先定義回應生成任務，接著介紹 COMMET 常識知識生成方法。

3.1 回應生成任務定義

此任務需要一個對話模型來扮演聽眾的角色並產生同理心的反應。輸入一個包含 N 個話語的對話歷史 $D = [u_1, u_2, u_3, ..., u_N]$，其中第 i 個話語 $u_i = [w_1^i, w_2^i, w_3^i, ..., w_{n_i}^i]$ 是由 n_i 個詞組成。我們的目標是產生一個與上下文一致、具有適當情感和信息豐富的回應 $Y = [y_1, y_2, y_3, ..., y_M]$。如圖1 所示，此爲所使用資料集中的對話資料，我們將 Dialogue History 的部分當作模型輸入，並希望模型經過訓練以後生成移情回覆。

3.2 常識知識獲取

考量到對話中會包含情緒以及信息，而 ConceptNet 較缺乏包含情緒的關係類別，ATOMIC 則在 xReact 這個關係類別中有較多的情緒字，故我們將使用 ATOMIC 作爲此兩個模型的常識知識庫。所提出的模型都是要以聽眾的角色來回覆說話者，所以在常識知識方面比較在意的是以說話者本人推斷的關係，在 ATOMIC 中對於事件和原因發生在參與事件的人身上總共有六種常識關係：事件對人的影響（xEffect）、人對事件的反應（xReact）、人在事件之前的意圖（xIntent）、爲了使事件發生人需要什麼（xNeed）、事件發生後人想要什麼（xWant），以及人的特徵屬性（xAttr）。由於 xAttr 這個類別是其他人推斷一個人的特徵，並不包含在移情反應中，因此我們使用的是剩餘的其他五個關係 (xEffect、xReact、xIntent、xNeed、xWant)。爲了達成給定事件生成常識推理，我們採用在 ATOMIC-2020 數據集上進行訓練的 COMET 模型去預測接下來回應中會含有的 Commonsense。

對於輸入歷史對話序列 D，分別將五個特殊 token ([xReact]、[xEffect]、[xWant]、[xNeed]、[xIntent]) 加在對話歷史中最後一個話語後面，並使用 COMET 針對每個關係生成五個常識知識，將生成的常識知識連接以獲得其常識序列 $K_r = k_1^r \oplus k_2^r \oplus ... \oplus k_5^r$，$r \in xReact$。

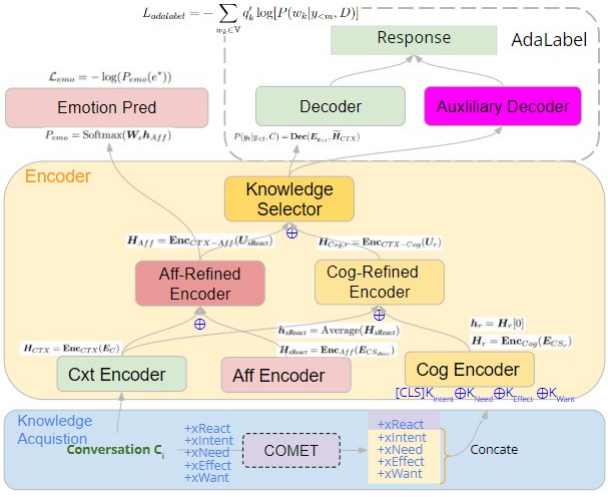

Figure 2: CEM-adalabel 架構圖

4 常識感知的移情回應生成模型

我們提出兩個增強生成多樣性的常識移情對話回應生成模型架構，分別是 CEM-AdaLabel 和 CE-BART。

4.1 CEM-AdaLabel

此模型基於 Transformer(Vaswani et al., 2017)，上下文及常識知識編碼器解碼器實現 CEM(Sabour et al., 2022)，而多樣性損失的部分實現 Adaptive Label Smoothing(AdaLabel)(Wang et al., 2021)，模型架構如圖2。由於論文空間限制，CEM-AdaLabel 方法請參考 (Huang, 2022)。

4.2 CE-BART

本論文則著重在介紹 Commonsense-aware Empathetic BART（簡稱 CE-BART）。此模型基於預訓練 BART 模型 (Lewis et al., 2020)，並結合在 ATOMIC-2020 數據集上進行訓練的 COMET 模型，模型架構如圖3，包含三個部分：常識知識獲取 (Knowledge Acquisition)、情感識別 (Auxiliary Emotion Recognition) 和回應生成 (Response generation)。

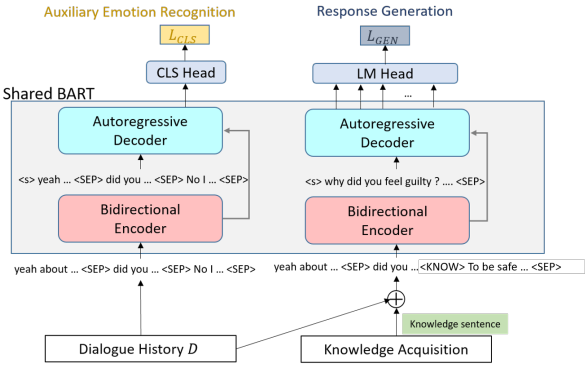

Figure 3: CE-BART 模型架構

BART 是一個基於 Transformer 架構的模型，有雙向編碼器 (Bidirectional encoder) 和自回歸解碼器 (Autoregressive decoder)，他的訓練方式是用任意噪聲函數破壞文本和學習模型來重建原始文本。與 BERT 不同的是，BERT 只使用 [MASK] token 去替換文本中的字，但 BART 爲了防止模型依賴像是序列長度的相關序列結構資訊，採用了多種 noise 函數去破壞掉這些資訊，所以它比 BERT 更適合自然語言生成的任務，且因爲包含雙向編碼器所以也比 GPT2 多了雙向上下文的信息。微調 BART 模型就可以快速地應用在其他任務上 (例如: 序列分類任務、序列生成任務、token 分類任務和機器翻譯)。而我們透過多任務學習 (Multi-task learning) 來學習回應生成和情感識別。

4.2.1 回應生成

將回應生成視爲序列生成任務，BART 模型本身有一個自回歸解碼器，所以它可以直接對序列生成任務進行微調。把對話歷史中的話語連接起來並在話語之間加上一個特殊 token <SEP>，而常識知識的部分是將章節3.2得到的 5 個常識知識序列連接並在前面添加特殊 token $[Know]$，目的是爲了讓模型知道這些是常識知識，最後將常識知識序列連接在對話歷史序列後面形成序列 $K = [K_{xReact} \oplus K_{xIntent} \oplus K_{xNeed} \oplus K_{xEffect} \oplus K_{xWant}]$，將此序列輸入到 BART 的共享嵌入層，來得到話語中每個 token 的隱藏狀態，然後將其送到 BART 的編碼器和解碼器，在訓練過程中解碼器的輸入會是右移 (right-shifted) 的回覆如圖3。目標回覆 $Y = [y_1, y_2, y_3, ..., y_M]$ 長度爲 M，生成回應損失 L_{GEN} 爲計算負對數似然損失 (Negative log-likelihood loss)。

$$L_{GEN} = -\sum_{j=1}^{M} \log P(y_j | D \oplus [Know] \oplus K, y_{<j}) \quad (1)$$

4.2.2 情感識別

情感識別可以被視爲序列分類任務，相同的輸入被送到 BART 編碼器和解碼器，然後取解碼器最後一個 token 對應的最終隱藏狀態作爲 label，輸入給一個線性多分類器 (multi-class linear classifier)。如果對話歷史 D 的正確情緒標籤是 e，則模型從 D 推斷出 e。一樣用負對數似然損失 (Negative log-likelihood loss) 計算分類損失 L_{CLS}。

$$L_{CLS} = -\log P(e|D) \quad (2)$$

4.2.3 Loss Weighting

模型訓練的損失 L 由兩部分組成：生成回應損失 L_{GEN} 和分類損失 L_{CLS}。L 是以上兩部分的加權和，它們的權重之和等於 1，α 是生成回應損失的權重。

$$L = (1 - \alpha)L_{CLS} + \alpha L_{GEN} \quad (3)$$

5 實驗

5.1 資料集

爲了驗證本篇論文所提出常識移情對話回應生成模型有加強生成多樣性的效能，在資料集上使用 ED (Rashkin et al., 2019) 來評估及比

較現有的方法。ED 是在 Amazon Mechanical Turk 上收集的大規模多輪移情對話資料集，包含約 25k 的一對一開放域對話，被廣泛使用於移情對話回應生成的基準資料集。收集方式是將兩個標記者配對：一個當作說話者、一個當作聆聽者。說話者被要求談個人的情感感受，聆聽者則是透過說話者所說的話推斷出潛在的情感，並做出善解人意的回應。該數據集提供了 32 個均勻分佈的情緒標籤。我們將對話歷史視爲模型輸入，將聆聽者的回應視爲目標輸出，整理後在訓練集中獲得 40,201 個對話，在驗證集中獲得 5,359 個對話，在測試集中獲得 4,836 個對話。

5.2 自動評估

我們採用 Perplexity (PPL) 和 Distinct-n (Dist-n) 作爲我們的主要自動評估指標。PPL 代表模型對其候選回應集的置信度，根據每個詞來估計一句話出現的概率，並用句子長度做正規化，如公式5 M 是句子長度，$p(w_i)$ 是第 i 個詞的概率。置信度越高，PPL 越低，可以用來評估生成回應的總體品質。

$$PPL = P(w_1 w_2 \dots w_M)^{-\frac{1}{M}} \tag{4}$$

$$= \sqrt[M]{\prod_{i=1}^{M} \frac{1}{P(w_i|w_1 w_2 \dots w_{i-1})}} \tag{5}$$

Distinct-n 測量生成的回應中不同 n-gram 的比例，公式6中，$Count(unique \quad n-gram)$ 表示回應中不重複的 n-gram 數量，$Count(word)$ 表示回應中 n-gram 詞語的總數量，Distinct-n 越大表示生成的多樣性越高，通常用於評估生成多樣性。此外，由於我們提出的模型將情感分類作爲訓練過程的一部分，因此也會評估情緒預測的準確度（Acc）如公式7，公式中 TP 是正確分類的正例數量，TN 是正確分類的負例數量，FP 是錯誤分類的負例數，FN 是被錯誤分類的正例數量。

$$Distinct\text{-}n = \frac{Count(unique \quad n\text{-}gram)}{Count(word)} \tag{6}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{7}$$

自動評估結果如表1，其中基線系統包括 MIME, KEMP, CEM 以及結合 CEM 與 AdaLable 的 CEM-AdaLabel (Huang, 2022)。我們提出的 CE-BART 上在所有自動評估指標方面都大大優於以上四種基線模型，CE-BART 在 PPL、Dist-1 和 Dist-2 上比沒有使用 COMET 生成常識知識資訊的 CE-BART 有更好的結果，代表增加外部知識有助於提高生成質量，儘管在情感準確性上有一點損失。實驗顯示 BART 模型因爲已經有在大規模的資料集訓練過，所以只要微調 BART 模型後，雖然在訓練上需要花得比較久，但得到的結果與沒有使用預訓練模型的方法在效能上有很大的差異。

5.3 Case Study

表2 列出了從基線 KEMP、CEM、BART 和我們提出的方法 CEM-AdaLabel 、CE-BART 生成的回應。在第一種範例中基線模型中，KEMP 將這句話語誤認爲是個好主意，而 CEM 生成的比較通用的話語去詢問說話者發生甚麼事了，CEM-AdaLabel 則是用偵測到說話者隱含的情緒 (scared) 所以知道有不好的事情發生，但還是缺少認知的常識知識。CE-BART 很好的偵測到說話者被嚇到並且因爲有人闖入他的家所以可能需要打電話給警察，相較於其他 Baseline 他生成包含情感 (oh no!) 和認知 (Did you call the police ?) 的回應。由此可知將預訓練模型與常識知識結合可以生成出更好的移情回覆。

第二個案例顯示了在多輪對話表達情感和認知的能力，KEMP、CEM 和 CEM-AdaLabel 忽略了說話者提到的"Thet've helped a a lot"，這句意味著他們的父母很好幫助了他們，所以 CE-BART 擷取到並認爲他們是很慷慨的很好的家人。

5.4 人工評估

在先前生成回應任務研究中，人工評估可以分成兩種方式進行。第一種是要求標註人員根據流暢性、相關性和同理心等方面對生成的回應進行 1 到 5 的評分；第二種是要求在同一對話歷史下的兩個模型之間選擇更好的回應。但是，給出 1 到 5 分的標準很可能在不同人之間有所不同，這導致標註者之間的一致性較低，所以此指標不適合評估模型性能。此外，要求標註者在沒有任何標準的情況下選擇更好的答案，僅依靠個人的喜好也不是可靠的指標，因爲每個人在做出選擇時可能會考慮不同的方面。

因此，我們參考 Sabour 等人 (Sabour et al., 2022) 的人工評估方法，對於給定的對話歷史，將 CEM-AdaLabel 和 CE-BART 的回應兩兩配對，並要求標註者根據對話歷史和以下三個標準選擇更好的回應：

Model	PPL	Dist-1	Dist-2	Acc(%)
MIME(Majumder et al., 2020b)	37.09	0.47	1.9	31.36
KEMP(Li et al., 2022)	36.89	0.55	2.29	**39.31**
CEM(Sabour et al., 2022)	36.11	0.66	2.99	39.11
CEM-AdaLabel	36.3	**0.79**	**4.08**	38.1
BART	12.87	1.9	9.11	**52.33**
CE-BART	**12.59**	**2.35**	**11.21**	51.29
Gold Response	-	5.98	36.84	-

Table 1: 自動評估的結果

Emotion	lonely
Dialog History	**Speaker:** I there, don’ t know what to do, just broke up with my girlfriend, we were 8 years together.
Knowledge	**sad**, to be alone, to break up with his girlfriend, **to find a new girlfriend**, gets depressed
KEMP	That is a good idea. I hope you have a great time!
CEM	Oh no! What happened?
BART	Oh no, that is terrible. Did you find out who did it?
CE-BART	Oh no , **i am sorry to hear that** . i hope you can **find a new girlfriend** soon .
Gold	Sorry to hear! Do you have any idea about the break up? Did you think about it?
Emotion	grateful
Dialog History	**Speaker:** When my wife and I had to move for her to go to school, we had a lot of trouble finding a place to live until my parents were nice enough to cosign on a loan. **Listener:** You must be so thankful ! **Speaker:** We are. They’ve helped a lot during my life, but this might have been the biggest one.
Knowledge	**happy**, **good**, to be grateful to them, **to be helpful**
KEMP	That is great to hear! I hope you have a great time.
CEM	That is great news!
BART	That is a great credit! I am glad you had a good support system.
CE-BART	That is so **generous** of them! You are so lucky to have such a **great family**.
Gold	They are really wonderful. Do you have plans to show how thankful you are?

Table 2: 本論文提出模型和 baseline 模型生成回應的案例研究

1. 同理心（Emp.）：哪個回應有很好的理解說話者的情況，且呈現出更貼切的情緒。

2. 連貫性（Coh.）：哪個回應更連貫並與對話歷史相關。

3. 信息量（Inf.）：哪個回應傳達了有關對話歷史的更多信息。

我們隨機從測試集抽取了 100 對回應，並分配了兩個人員來標註每一對。標記允許平手，但鼓勵標註者盡量選擇其中一個回應，並使用 Cohen’s kappa 係數 (κ) 來分析兩個標註者之間的一致性，其中 $0.4 < \kappa < 0.6$ 表示中等一致性。

如表3 所示，CE-BART 在三個方面都優於 CEM-AdaLabel ，所以使用預訓練模型後能夠產生更連貫、包含同理心和更多信息的回應。在連貫性的部分 CEM-AdaLabel 相較於其他兩個部分比例稍高的原因，是因爲我們觀察到 CE-BART 會生成較長的回應 (13.23 個單詞/回應)，而 CEM-AdaLabel 會生成較短的回應 (10.24 個單詞/回應)。造成標註者認爲 CE-BART 生成的回復可能只有前半部分是與對話歷史相關，而 CEM-AdaLabel 較短但整句是與對話歷史有關的。

6 結論

現有教育型對話機器人並未結合移情對話生成模組，使得機器人只有講故事或問問題的功能。我們引用了「動態回顧循環」（Active

Comparisons	Aspects	Win	Loss	κ
CE-BART vs. CEM-AdaLabel	Emp.	86%	5.5%	0.61
	Coh.	85.5%	9%	0.73
	Inf.	94%	2.5%	0.56

Table 3: 人工評估的結果

Reviewing Cycle）提問法，透過 4F 解說技巧—事實（Facts）、感受（Feelings）、發現（Findings）及未來（Future），引導學生從不同角度反思時故事內容、個人感受、成長經驗、以及未來的發展方向，而透過 Feeling 和 Future 講到個人日常經驗的內容時，尤其需要結合常識知識圖譜來增強機器人的回應多樣性。我們提出利用預訓練的 BART 模型結合 COMET 生成的常識知識來生成移情對話，雖然在訓練時間上需要 3 個小時多，但在自動評估、案例研究和人工評估都表明，我們提出的 CE-BART 都優於 MIME, KEMP, CEM, 及 CEM-AdaLabel 等基線模型，並證明了移情對話的生成受益於預訓練模型和外部知識。

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787—2795, Red Hook, NY, USA. Curran Associates Inc.

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

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Tzu-Hsien Huang. 2022. Two simple ways to improve commonsense-aware empathetic response generation. Master's thesis, National Central University.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pengfei Li, Peixiang Zhong, Kezhi Mao, Dongzhe Wang, Xuefeng Yang, Yunfeng Liu, Jianxiong Yin, and Simon See. 2021. ACT: an attentive convolutional transformer for efficient text classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13261–13269. AAAI Press.

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

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.

Chen-Chung Liu, Mo-Gang Liao, Chia-Hui Chang, and Hung-Ming Lin. 2022. An analysis of children' interaction with an ai chatbot and its impact on their interest in reading. *Computers & Education*, 189:104576.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020a. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020b. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*, pages 8968–8979.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

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*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444—4451. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying dialog generation via adaptive label smoothing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. 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 1571–1582, Online. Association for Computational Linguistics.