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

應用自動資訊擷取於故事書問答生成之研究
Applying Information Extraction to Storybook Question and Answer Generation

高愷言 **Kai-Yen Kao**
中央大學資訊工程學系
kykao@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

從故事文本中產生高品質且通順的問題—答案配對是一件耗時且耗力的事情，問題的生成目的不是要讓學生回答不出來，而是幫助學生了解故事所要傳達的訊息，因此需要經過巧妙的設計將文本中的重要資訊當成答案，並且生成與之相對應的問題。在本文中，我們通過將問題類型及其類型定義結合到輸入中來改進 FairyTaleQA 問題生成方法，以微調 BART (Lewis et al., 2020) 模型改進問題生成效能。此外，我們進一步利用 (Zhong and Chen, 2021) 中的實體和關係提取作爲基於模板的問題生成的元素。使用 pipeline 的方法 (Zhong and Chen, 2021)，最後將擷取出來的關係作爲模板式生成的要素。

Abstract

For educators, how to generate high quality question-answer pairs from story text is a time-consuming and labor-intensive task. The purpose is not to make students unable to answer, but to ensure that students understand the story text through the generated question-answer pairs. In this paper, we improve the FairyTaleQA question generation method by incorporating question type and its definition to the input for fine-tuning the BART (Lewis et al., 2020) model. Furthermore, we make use of the entity and relation extraction from (Zhong and Chen, 2021) as an element of template-based question generation.

關鍵字：問題-答案配對生成、問題回答、資訊擷取

Keywords: Question-Answer Pairs Generation, Question Answering, Information Extraction

1 簡介

大量閱讀是語文教育相當重要的一環，可以讓學生在學習中保持新鮮感與熱情。若能搭配互動式教學，與學生進行故事書討論，藉由提出故事書中的問題，讓學生思考並且回答，可以評估學生對於故事書的理解，同時增進學生口語表達能力。本篇研究的目標是幫助老師自動生成與故事相關的問題—答案配對，讓學生能根據問題去思考閱讀過的故事書，培養閱讀以及統整資訊的能力。

Xu 等人 (Xu et al., 2022) 將故事書的提問方式分爲 7 類，包括角色 (Character)、場景 (Setting)、發生事件 (Action)、人物感受 (Feeling)、因果關係 (Causal Relationship)、產生結果 (Outcome resolution)、以及未來預測 (Prediction)，透過這七種面向來設計問題—答案，涵蓋故事文本中大部分的內容。

問題生成的方式主要有兩種，模板式產生以及使用序列模型進行生成，傳統的模板式生成大多是以人工產生的規則進行問題的產生，本研究利用實體與關係擷取技術，可以將文本中重要的資訊擷取出來，取代人工撰寫模板的複雜成本，使得模板式生成也可以產生出具有品質的問題—答案配對。生成式的模型，可以取決於是否先提供答案進行生成，本研究使用先提供答案 (answer-aware) 的方式來生成問題，擷取答案的方法可以使用 Heuristic-based 的方式，故只要使用一個模型就可以生成問題—答案配對。

在本篇論文中，我們對於問題生成使用了兩種方法，在生成式的方法中，我們加入了問題類別與其定義，在 FairyTaleQA(Xu et al., 2022) 的資料集中與 baseline 模型相比可以在 $ROUGE-L$(Recall-Oriented Understudy for Gisting Evaluation) 上提升 0.034，接著我們在生命教育故事進行人工評估，加入問題類別與其定義的問題皆獲得比較高的分數，藉由自動評估與人工評估，皆顯示加入問題類別與定義可以獲得更好的成果。我們也透過了問題回答的任務來對於問題—答案配對進行評估，其結果與人工評估之間的相關性是比使用 ranking model 的相關性略高，也說明了我們使用這種評估方法也可以很好的對於問題—答

案配對進行篩選。在模板式生成問題方面，我們使用了自動資訊擷取的技術，將故事中文本的資訊擷取出來，應用在模板式問題生成，可以對於問題生成有另一種角度的產生。

2 相關研究

2.1 問題生成 (Question Generation)

問題生成 (Question Generation,QG) 是給定一段文本或句子，生成具有可讀性以及與文本相關的流暢問句。傳統的問題生成作法是屬於規則式的產生 (Das et al., 2016)，藉由人工設計的規則、模板、句法分析將文本中的句子轉變爲疑問句的形式，以此來獲得問題，但是這樣的作法會因爲文本的豐富度大量耗費人力，擴展性也不佳，後續較少使用。

問題生成也可以用是否給定「答案」來做區分。有答案 (answer-aware) 的問題生成 (Zhou et al., 2017) 是會先根據文本內容產生特定答案，這個答案可以是文本中的詞語、句子或是人工產生的，在生成問題時，答案可以提供模型在生成問題時有更多的資訊，也可以將問題限制在答案下生成。無答案 (answer-unaware) 的問題生成顧名思義在生成時沒有給定答案 (Du et al., 2017)，模型可以對文本中的任意位置進行問題的產生，沒有答案的資訊以及約束，其產生的問題會較分散且參差不齊，這方面的研究相對 answer-aware 少很多。

Zhou 等人 (Zhou et al., 2017) 在 seq2seq(Sequence to sequence) 架構下，增加了答案 (answer-aware) 進行生成，後續對於問題生成的任務多採用 seq2seq 的架構來實做，而預訓練模型的出現也使得這個任務又有更好的成效，故現在的研究幾乎都使用預訓練模型：而使用 BERT(Dai et al., 2019)、BART(Lewis et al., 2020)、T5(Raffel et al., 2020) 等。

Xu 等人在 2022 年提出的 FairyTaleQA(Xu et al., 2022) 的資料集是以故事爲目標所建立的問答生成資料集，其故事來源是來自古騰堡計畫 (Project Gutenberg) 所收集的書籍，並且以"Fairytale" 爲關鍵字所蒐集的故事書共 278 本。本研究參考 Paris(Paris and Paris, 2003) 所提出的 7 種類型問題，邀請教育專家根據故事文本中的內容、以及 7 種類別，進行問題—答案配對的生成。

2.2 資訊擷取 (Information Extraction)

資訊擷取 (Information Extraction)，是從自然語言文本 (非結構性資料) 中，抽取結構化資料的一個過程，是自然語言理解的基本任務。主要可以分爲三個子任務：實體擷取 (Entity Extraction)、關係擷取 (Relation Extraction)、事件擷取 (Event Extraction)。

2.2.1 實體擷取 (Entity Extraction)

實體擷取 (Entity Extraction, EE) 是一種從文本中，將命名實 lewis-etal-2020-bart 體識別、擷取爲預定義類別的任務，包括：人物 (PER)、地點 (LOC)、組織 (ORG)、時間 (TIME) 等，將原本非結構性的文本，進行序列標記，擷取出特定實體。

實體擷取的主流方法大概可以分爲兩種，較爲傳統的作法爲透過建立辭典 (Wu et al., 2020)，再用其來比對文本中的詞語，另一種是透過機器學習訓練模型進行序列標記的任務來擷取出實體。

在進行實體擷取時，常常會結合 CRF(Conditional Random Field)，將 CRF 當作模型的輸出層。而雙向長短期記憶模型 (Bidirectional Long Short-Term Memory，BiLSTM)，BiLSTM-CRF(Huang et al., 2015) 的架構可以利用記憶功能來保留較長距離的上下文資訊，以此來提升序列標記的準確度。而使用 Bidirectional Encoder Representations from Transformers (BERT)(Dai et al., 2019) 對於實體擷取的任務又能更上一層樓，將 BERT 的預訓練模型當作 embedding 層，除了可以更好的保留上下文關係，更可以有效提升小樣本資料的擷取準確度。

2.2.2 關係擷取 (Relation Extraction)

關係擷取 (Relation Extraction, RE) 是資訊擷取 (Information Extraction, IE) 中很重要的子任務，其可以將文本中一對實體之間的語義關係所擷取出來，大多數的關係擷取都是以二元關係爲主，關係可以定義爲 (e_i, r, e_j)，e_i 與 e_j 分別代表兩個實體，r 代表兩個實體之間的關係，可以把原本非結構性的文本內容整理成結構化且具有意義的資訊。

關係擷取通常會搭配實體擷取任務一起進行，目前主流的方法有兩種 (Zhong and Chen, 2021)，第一種是 Pipeline 的方式，將兩者視爲獨立的任務，兩個任務的模型訓練不會互相影響，另一種方法是將兩個任務進行聯合學習 (Joint Learning)，其優點是兩個子任務之間的資訊可以共同被使用，用來協助另一個子任務進行預測，缺點則是整體模型的架構龐大且複雜。

Pipeline 的方法裡，Zhong 等人 (Zhong and Chen, 2021) 在進行實體與關係擷取時採用了 pipeline 的架構，這個模型是本篇論文在做實體與關係擷取所使用的模型。

在聯合學習的方法中，Shang 等人 (Shang

et al., 2022) 提出了一種將關係擷取視爲分類任務的方法，而這篇論文所提出的標記方法也可以有效的降低在標記時所浪費的空間以及時間成本。

2.2.3 事件擷取 (Event Extraction)

事件擷取 (Event Extraction, EE) 是一種從非結構化的文本中，擷取出與目標相關的事件與相關資訊，識別特定類型的事件後，並將事件中的要素標示出來。根據任務的需求，事先訂定事件的類別，擷取的資訊大致可以分爲以下幾種：事件類別 (event type)、觸發詞 (trigger word)、事件要素 (event arguments)。

事件擷取的流程大致可以分爲以下幾個部分：事件觸發詞偵測、事件觸發詞分類、事件要素偵測、事件要素分類，早期的研究可以分爲 Pipeline 與聯合學習兩種，前者的缺點是會有錯誤傳遞的問題，可以利用聯合學習的方式來改善，而不管是哪種方法，都需要大量的標記資料，也需要設計各個子任務的最佳組合，因此 End2End 生成式的事件擷取漸漸被提出來作爲選項之一。

因此 Lu 等人 (Lu et al., 2021) 提出 Text2Event 模型，是一種 Sequence-to-Structure 的事件擷取方法，採用 End2End 的方式直接從文本中擷取事件。

3 Generative 問題生成方法

生成式 (Generative based) 的問題產生 (Question Generation) 一般採用 pipeline 的架構，並使用 answer-aware 的方式，以先擷取的答案來當作生成問題的參考，最後透過 Ranking model 與問題回答 (Question Answering) 來進行評估。這個任務是基於預訓練模型 BART(Lewis et al., 2020) 進行實作，以下依序介紹問題定義、模型架構、資料集、實驗。

3.1 問題定義

在問題產生任務中，會先將文本分成多個句子：$S_1, S_2, ...S_N$，根據句子 X_i 產生跟語句相關的問題 $q_i^1, q_i^2, ...q_i^n$ 與答案 $a_i^1, a_i^2, ...a_i^n$，形成問題—答案配對 (q_i^j, a_i^j)，這個任務的輸出就是多組的 $(q_i^j, a_i^j), 1 < i < N$，最後並爲每個問題—答案配對產生分數。

3.2 模型架構

我們參考 Yao 等人 (Yao et al., 2022) 提出的架構，採用 pipeline 的方式進行問題生成，如圖1所示，架構包含三個 module，第一個爲答案產生模組，第二個爲問題產生模組，第三個爲問題—答案配對排序模組。由於這個架構是在 answer-aware 的情況下產生問題，可以得到與答案相對應的問題，也可以根據答案的種類，生成不同種類的問題，讓整體的問題—答案配對品質更好，故先利用 heuristics-based 的答案產生模組，先產生答案來當作問題生成的指引。

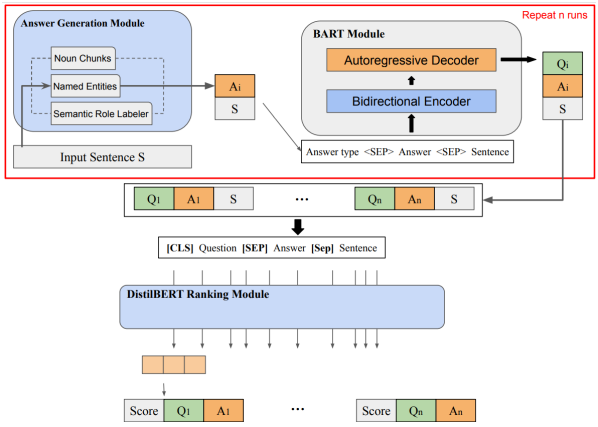

Figure 1: 問題生成模型

Yao 等人對於答案產生模組的做法是透過 Spacy(Honnibal and Montani, 2017) 的套件來擷取出命名實體與名詞片語，並利用 AllenNLP(Gardner et al., 2018) 的 bert-base srl 來進行語義角色標註 (Semantic Role Labeling)，將句子以動詞進行切分，可以將與動詞相關的主詞、受詞解析出來，最後組成主詞、動詞、受詞來當成候選的答案。

第二個模組是問題產生的核心，由於 BART 在預訓練時就是使用 sequence-to-sequence 的方法進行訓練，所以很適合用來進行序列產生的下游任務 fine-tuned。不同於 Yao 等人的做法，我們除了使用答案以及句子外，還加入了答案的類別以及此類別 (角色、場景、感受、動作、因果、結果、預測) 的定義 (如表1) 進行訓練，讓問題可以根據類別生成更適合的問題。由於 BART 本身具有 Autoregressive Decoder，所以可以直接對於生成任務進行微調，將 source：問題類別、答案、句子中間分別以 <SEP> token 連接起來輸入 Encoder，再將 target：根據答案與句子生成的問題輸入 Decoder，即可進行訓練。在進行問題生成時，輸入問題類別、答案、句子，模型即可生成出相對應的問題，來完成得到問題—答案配對的任務。

第三個模組是使用 DistilBERT(Sanh et al., 2019) 來對於經過前面兩個模組產生的問題—答案配對進行排序，可以將排序的任務視爲模型產生的問題—答案配對與訓練資料標記的問題—答案配對的分類任務，透過 DistilBERT 進行下游任務：序列分類 (Sequence Classification) 的微調，在進行分類任務時，需要再

Question-Answer Type	Definition
Character	Ask test takers to identify the character of the story or describe characteristics of characters.
Setting	Ask about a place or time where/when story events take place and typically start with "Where" or "When."
Feeling	Ask about the character's emotional status or reaction to certain events and are typically worded as "How did/does/do . . . feel"
Action	Ask characters' behaviors or additional information about that behavior
Casual Relationship	Focus on two events that are causally related where the prior events have to causally lead to the latter event in the question. This type of question usually begins with "Why" or "What made/makes."
Outcome Resolution	Ask for identifying outcome events that are causally led to by the prior event in the question. This type of question is usually worded as "What happened/happens/has happened. . . after..."
Prediction	Ask for the unknown outcome of a focal event. This outcome is predictable based on the existing information in the text

Table 1: FairytaleQA 7 種問題—答案類別 (Xu et al., 2022)

輸入句子的最前端加入特殊 token [CLS] 來代表模型是要進行分類任務，最後根據模型產生的標籤機率，轉換爲分數後，即可得到問題—答案配對的分數。

3.3 實驗

我們採用將 FairytaleQA(Xu et al., 2022) 資料集做爲我們實驗的資料集。其故事書共 278 本，訓練資料集包括 232 本書、8548 個 QA-pairs；驗證資料集包括 23 本書、1025 個 QA-pairs；測試資料集包括 23 本書、1007 個 QA-apirs。這些問題—答案配對的比例如圖2，大體上 7 個類別在訓練、驗證、測試集的比例皆算一致。

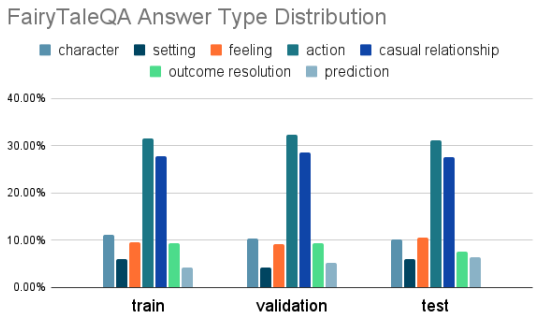

Figure 2: FairyTaleQA 答案類型分佈

3.3.1 評估方法

模型生成出來的問題可以使用 *ROUGE-L* 指標來和標記資料進行評估，其中 L 代表最長公共子序列 (Longest Common Subsequence, LCS)，其中 X 爲 golden sequence，Y 爲生成的句子，見公式1。

$$ROUGE\text{-}L = \frac{LCS(X,Y)}{len(X)} \quad (1)$$

3.3.2 Testing on Given Answers

我們首先忽略答案產生模組的影響，利用測試資料中給定的問題答案配對中的答案做爲輸入，探討使用不同輸入 fine-tuned 的模型在 FairytaleQA 測試資料上面的效能，以及加入問題種類對於整個模型生成問題的效果有所提升。由於每一個 QA 配對均有問題類型，我們在訓練時可以加入問題類型以及定義來進行訓練。測試時，我們再將 QA 配對中的答案當成模型的輸入，再將產生的問題與 QA 配對中的問題進行 $ROUGE-L$ 評估，在 $ROUGE-L$ 約提升 0.2，若是更進一步加入問題答案類別的定義，模型可以在效能上再取得 0.16 的進步，在生成時可以更加根據問題類別來產生相對應的問題。

經過排序模組後，我們取評估分數爲正的問題做平均，以及計算評估分數爲正的問題—答案配對數量。見表2，加入問題種類進行 fine-funed 的模型，可以產生更多數量 (676 –> 690) 的問題、以及更高的分數 (6.30–>6.36)。

3.3.3 Testing on Heuristic Answers

確認使用問題種類可以生成更好的問題後，接著進行由答案產生模組輸出的答案做爲問題生成的輸入，答案產生方法分爲 Entity、Noun Chunk、Semantic Role Labeling (SRL)3 類，其中 entity 的平均長度爲 1.66、chunk 的平均長度爲 2.46、SRL 的平均長度爲 9.46。

以下比較不同答案產生方式在生成問題上的效能比較。見表3，我們選擇 ranking score 爲正的分數進行比較，並且計算其數量。entity 的平均分數是小幅領先 srl，與 chunk 之間

	ROUGE-L	Mean Positive Score	Count
Baseline(Yao et al., 2022)	0.506	6.30	676
+ question type	0.524	6.36	690
+ question type and definition	**0.540**	**6.37**	**692**

Table 2: 生成式問題於 fine-tuned 模型的效能

有著 0.65 的差距，與給定的答案進行實驗時一樣，在加入問題類別與其定義後，可以在 ranking score 取得較好的表現。

3.3.4 Testing on 24 Life Educational Story Books

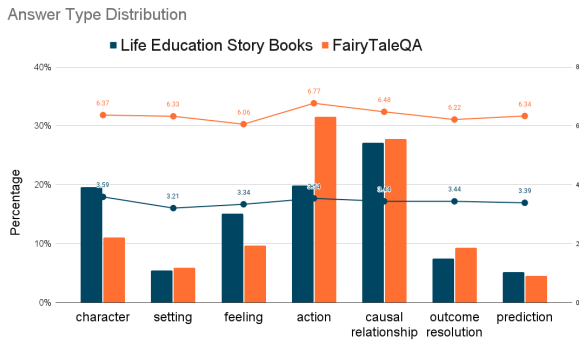

Figure 3: FairyTaleQA 與生命教育故事問題類別分布

最後在 24 本生命教育故事書來測試模型，先採用 Heuristic-based 的方法擷取出故事文本的答案，在透過 3 種不同的輸入方式來產生問題—答案配對，結果如上表4，由結果可以看到，在不同的答案上，有加上問題種類與其定義來進行生成在自動評估上是可以得到比較好的結果。而利用實體擷取模型擷取出來的 Entity 也可以在 ranking score 上面幫助模型生成問題，不管是在數量與 Ranking Score 皆獲得比較好的表現，見表5。

接著比較 FairytaleQA 測試集 (人工標記問答) 與生命教育故事的問題類別分布，見圖3，左邊的 Y 軸是表示不同問題類別的百分比，可以看到兩者的比例是相當接近的，也就是在生命教育故事來進行後續的實驗是相當洽當的。右邊的 Y 軸則是表示各個類別的 Ranking Score，上方的折線分別表示兩個資料集在 7 種問題類別的 ranking score。

這邊探討 7 種不同的問題類別分數的分布，如圖4，可以看到在不同類別的分數主要集中在中間，呈現一個類似於常態分布的狀態。

3.3.5 使用問題回答進行評估

問題回答的模型架構基本上與 BART-Based 的問題生成相同，都是 sequence-to-sequence 的架構，差別只在於輸入與輸出的差異，對

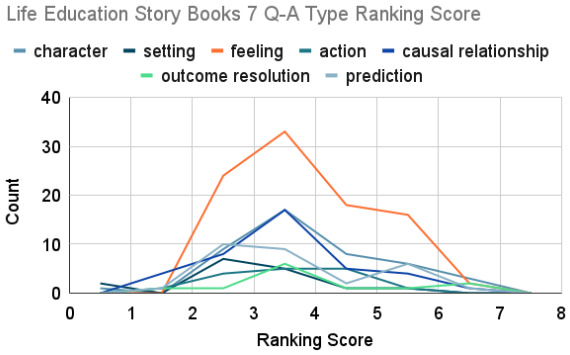

Figure 4: 生命教育故事問題類別分數分布

於問題回答任務，其輸入是以問題、文本中間加 <SEP> token 做爲 Encoder 輸入，透過 Decoder 輸出答案進行訓練，在生成時模型即可以根據問題與文本找出相對應的答案。

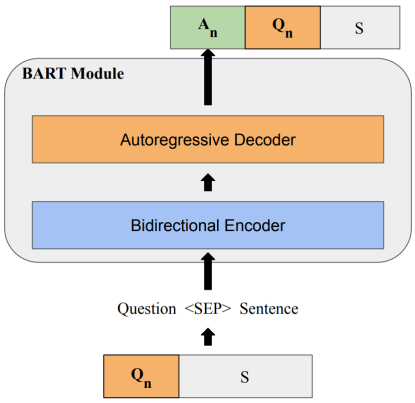

Figure 5: BART-Based 問題回答模型

我們使用 FairyTaleQA 以及 SQuAD 兩個資料集進行訓練，而測試的資料集則選擇兩個，第一個是 FairyTaleQA 的測試資料集中的人工標記問答以及 Heuristic Answer，第二個是 24 本生命教育故事書經過問題生成後的問題—答案配對進行評估。透過 Rouge-L 來計算生成的答案與標記答案之間的相似度。如表6，可以看到使用 SQuAD 進行訓練比 FairyTaleQA 的效果還要好，其因是 SQuAD 的資料數量是 FairyTaleQA 的將近 10 倍，故有更好的效能。

透過問題回答這個任務，我們可以用來評估問題—答案配對生成的品質，提供除了排序

Mean Positive Score	Entity	Noun Chunk	SRL	Avg.	Count
Baseline (Yao et al., 2022)	4.27	3.84	4.08	4.06	40
+ question type	4.47	4.00	4.35	4.27	56
+ question type and definition	**4.66**	**4.01**	**4.65**	**4.44**	**70**

Table 3: Heuristic Answer 於 fine-tuned 模型的效能：Mean Positive Score 和 Count

Models	Mean Positive Score		Count	
	Chunk	SRL	Chunk	SRL
Baseline (Yao et al., 2022)	2.83	4.03	5	17
+ question type	3.10	3.92	6	29
+ question type and definition	**3.23**	**3.98**	**10**	**29**

Table 4: 24 本生命故事於 fine-tuned 模型的效能 (Chunk, SRL)：Mean Positive Score 和 Count

Models	Mean Positive Score		Count	
	Entity	ACE Entity	Entity	ACE Entity
Baseline (Yao et al., 2022)	3.43	3.22	41	107
+ question type	3.45	3.51	58	131
+ question type and definition	**3.69**	**3.75**	49	116

Table 5: 24 本生命故事於 fine-tuned 模型的效能 (Entity, ACE Entity)：Mean Positive Score 和 Count

Rouge-L	FairyTaleQA	SQuAD
FairyTaleQA (Gold Answer)	0.491	**0.502**
FairyTaleQA (Heuristic Answer)	0.445	**0.484**
Life Education Story Books	0.474	**0.675**

Table 6: 問題回答效能比較: ROUGE-L

模組第二個評估方法。換言之，我們會計算問題—答案配對中的答案與問題回答所生成的回答之間的最長公共子序列，以 Rouge-L 做爲評估的指標。圖6是利用問題回答來對 24 本生命教育故事書產生的問題答案配對進行評估，可以看到在 feeling、causal relationship 以及 outcome resolution 的地方，srl 是可以比 entity 與 chunk 得到更好的結果，因爲其擷取出來的答案是具有動詞以及相關資訊，較可以用來提供這方面的問題。

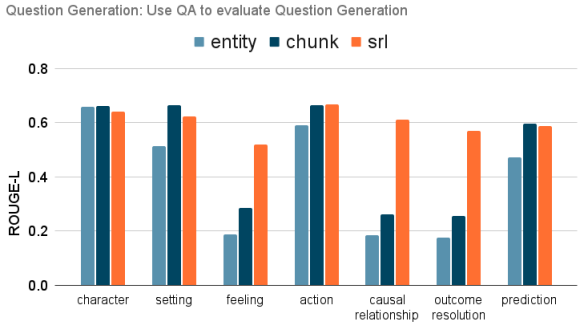

Figure 6: 使用問題回答對 24 本生命教育故事進行評估 (Rouge-L)

圖7是使用散佈圖來觀察使用問題回答以及 ranking model 兩種不同方式評估之間的相關性，可以看到兩者之間的分布較散，有些 Rouge-L 分數爲 0 的問答句的分數也相當高，顯然並不合理。因此我們進一步採用人工評估對生成的問答配對進行評估。

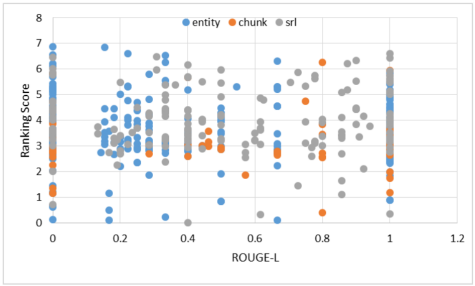

Figure 7: 使用問題回答對 24 本生命教育故事進行評估散佈圖

3.4 人工評估

由於在做序列生成任務時，除了上述自動評估的方法可以當成效能的指標，問題的通順程度以及問題—答案配對之間的相關性是較難衡量的，這時候就需要人工評估來協助判斷，這邊以 3 個指標來對於問題—答案配對進行評估。第一個指標是問題的通順程度 (Question Readability)，這個指標可以用來檢視所生成的問題是否通順，符合閱讀的直覺，第二個指標是問題與文本之間的相關程度 (Question-Text Relevancy)，用來判斷問題是否有問到文本中的內容，第三個指標是答案的相關性

(Question-Answer Relevancy)，用來說明問題—答案配對之間的相關程度，用來判斷所產生的問題與答案之間是否能有很好的匹配性。每個指標的分數區間爲 0 到 5，進行人工評估的標註者爲 3 人，最後以三人平均分數呈現。

首先使用 24 本生命故事書，每個問題類別各隨機挑選一個問答配對（不排除負值的排序分數），由三位研究生進行人工評估。結果如圖8所示，可以看到各個類別在不同評估指標下的分數差異。

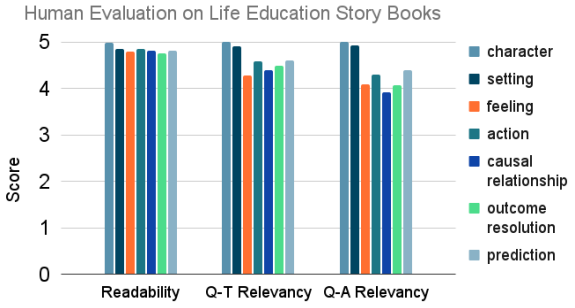

Figure 8: 人工評估結果：7 種問題類別

從人工評估與排序分數散佈圖 (圖9) 可以看到 Ranking module 在分數上分布的較不均勻，分數從高到低皆有，相關係數僅爲 0.121。而人工評估與 Rouge-L 的散佈圖 (圖10) 可以看到右上角分布的較有一致趨勢，相關係數爲 0.245，顯示透過問題回答來進行問題—答案配對的評估比較有意義。

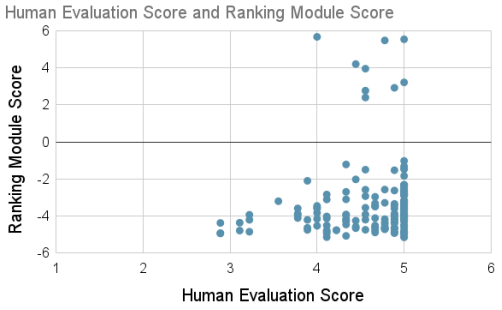

Figure 9: 人工評估結果與 Ranking Score 散佈圖

接著使用 24 本生命故事來進行人工評估的資料，在當中選取 10 本故事書，每本故事書取 3 組問題—答案配對來進行，分別比較不同的模型輸入的影響，共產生 90 組 QA pairs ，結果見表7前三列。由人工評估的結果來看，與自動評估相同，不管是在語句的通順程度抑或是問題—答案配對之間的相關性，加上問題種類與其定義確實能夠在問題生成上得到更好的結果。

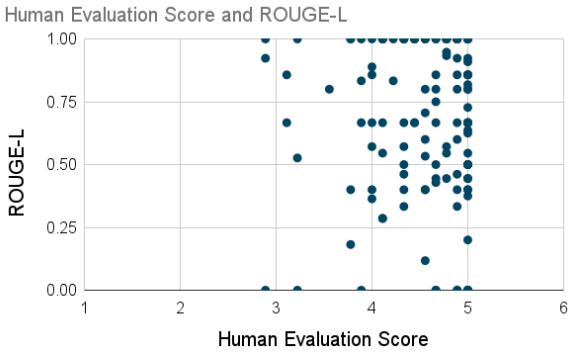

Figure 10: 人工評估結果與問題回答評估分數散佈圖

3.5 Case Study

在給定相同的故事內容以及答案的情況下，加入問題類別以及定義的問題生成可以產生具有更多資訊的語句，如表8、9，在生成問題時，原本的 baseline 模型生成的問句較爲粗略，問得比較是大範圍的問題，而加入問題類別以及定義的生成問句，則是可以看出來有提到文章中的更多資訊，描述的也比較精準。

4 Extraction Enhanced 問題生成

在這個章節，會介紹如何使用資訊擷取任務來幫助模板式的問題生成，其中分爲兩個部分：關係 (Relation) 與事件 (Event)，最後在對模板式的問題生成進行人工評估。

4.1 模板式問題生成方法：Relation

故事文本經過關係模型擷取 (Zhong and Chen, 2021) 後，會得到兩個實體之間的關係。透過這種結構，可以設計模板式的問題產生，使用關係類別來當作答案的部分，而問題產生可以是詢問「兩個實體之間是什麼關係?」，以此得到問題—答案配對。抑或是與 7 種答案類別進行配對，以此生成多樣性的問題。關係擷取的部分，共有 6 種類別以及 18 種子類別，如表10：

Type	Subtype
ART	User-Owner-Inventor-Manufacturer
GEN-AFF	Citizen-Resident-Religion-Ethnicity, Org-Location
ORG-AFF	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE	Artifact, Geographical, Subsidiary
PER-SOC	Business, Family, Lasting-Personal
PHYS	Located, Near

Table 10: ACE2005: Relation Types (Walker et al., 2006)

	Question Readability	Q-T Relevancy	Q-A Relevancy
Baseline (Yao et al., 2022)	4.60	3.87	4.06
+ question type	4.52	4.07	4.27
+ question type and definition	**4.84**	**4.64**	**4.61**
Template-based	**4.96**	4.60	3.96

Table 7: 人工評估結果

Story Text	All the notes in her music books and all the things related to music were all gone.
Answer	all the things related to music
Baseline (Yao et al., 2022)	what were all gone from the house?
+ question type	what was missing from the girl's books?
+ question type and definition	what was missing from the notes in her music books?

Table 8: 24 本生命故事：Question Generation Example 1

Story Text	"Yes dear. That's because Sophie is very special. She has Down Syndrome," her mom explained.
Answer	Down Syndrome
Baseline (Yao et al., 2022)	what was special about Sophie ?
+ question type	what was special about Sophie ?
+ question type and definition	what kind of special condition does Sophie have ?

Table 9: 24 本生命故事：Question Generation Example 2

4.2 模板式問題生成方法：**Event**

使用 text2event(Lu et al., 2021) 的 API 可以從故事文本中擷取出事件，其中包含三個部分，第一部分爲 *Role*：代表這個事件中的人、事、時、地、物，第二部分 *Type*：代表事件的類別，如攻擊、結婚、運輸... 等，第三部分爲 *Trigger*：代表句子中被判斷爲事件的關鍵字。

對於事件的模板式問題生成，可以一樣分爲三個部份去詢問，從 *Role*、*Type*、*Trigger* 的角度來產生問題，以這幾種模板式的方式可以增加問題—答案配對的豐富性。同樣的可以用 7 種答案類別來生成模板性的問題。事件擷取的部分，共有 8 種類別以及 33 種子類別，如表11：

Type	Subtype
Life	Be-Born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-Ownership, Transfer-Money
Business	Start-Org, Merge-Org, Declare-Bankruptcy End-Org
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
Personnel	Start-Position, End-Position, Nominate, Elect
Justice	Arrest-Jail, Release-Parole, Trial-Hearing Charge-Indict, Sue, Convict, Sentence, Fine Execute, Extradite, Acquit, Appeal, Pardon

Table 11: ACE2005: Event Types (Walker et al., 2006)

4.3 評估

在人工評估的部分，模板式的問題在問題的可讀性上取得了很高的分數，見表7的最後一列，因爲是透過模板來產生問題，所以可以取得通順的問題，而在問題—答案之間的關聯性，則是因爲擷取出來的資訊是固定模板的，故在整體關聯性上會較低，此時就可以透過生成式的問題—答案配對來補足。

5 結論

本研究主要在探討如何從故事書中產生高品質且具有多樣性的問題，在生成式模型的部分，採用先擷取出答案，再根據答案來產生與之相對應的問題，這種方法可以有效的將問題與答案之間的相關性提高。透過加入問題類別與定義，提升問題—答案配對的生成。在模板式問題生成的部分，我們採用了實體與關係擷取的結果，應用於樣板式問答生成，搭配 7 種問題類別來產生不同面向的問題。

最後我們使用問題回答的方式來對於問題—答案配對進行 Rouge-L 評估，相較於排序模型，Rouge-L 與人工評估的相關性更高。透過排序模型及問題回答這兩種評估方法，更有效的篩選生成的問題—答案配對。

References

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5.

Rubel Das, Antariksha Ray, Souvik Mondal, and Dipankar Das. 2016. A rule based question generation framework to deal with simple and complex sentences. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 542–548. IEEE.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991.

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.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11285–11293.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. Named entity recognition with context-aware dictionary knowledge. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 915–926, Haikou, China. Chinese Information Processing Society of China.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *North American Association for Computational Linguistics (NAACL)*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer.