CDGP: 基於語言預訓練模型之克漏字干擾選項自動生成研究

CDGP: Automatic Cloze Distractor Generation based on Pre-trained Language Model

組員：江尚軒、王思正

指導教授：范耀中 教授

1. 前言
2. Candidate Set Generator (CSG)
3. Distractor Selector (DS)
4. 相關實驗
5. 結論
6. 參考文獻
7. Q&A

我們做了什麼？

- 我們提出的框架**CDGP**達成了**自動化生成克漏字干擾選項**的目標，並首次提出基於**深度學習**之**語言預訓練模型**的方法。

- 經實驗證實，CDGP已經**超越AAAI2021研究[1]所展示之效能，顯示我們為目前克漏字干擾選項生成最佳效能之方法。**

- 也經由**人工測驗**的方式，證實可以實際應用於克漏字出題。

[1] Ren, S., & Q. Zhu, K. (2021). Knowledge-Driven Distractor Generation for Cloze-Style Multiple Choice Questions. Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4339-4347

克漏字問題是？

- 克漏字問題組成為一段有挖空的**考題**，一個適合填入挖空的**答案**以及幾個誤導測試者的**干擾選項**。

考題	If you want recovery soon, start by feeling grateful that you are still ____.
選項	(A) alive 答案 (B) lovely (C) lively (D) living 干擾選項

干擾選項生成困難點

- 以人的角度來看，生成克漏字的干擾選項並非困難的事。**然而對於機器來說，卻不是一件容易的事**。

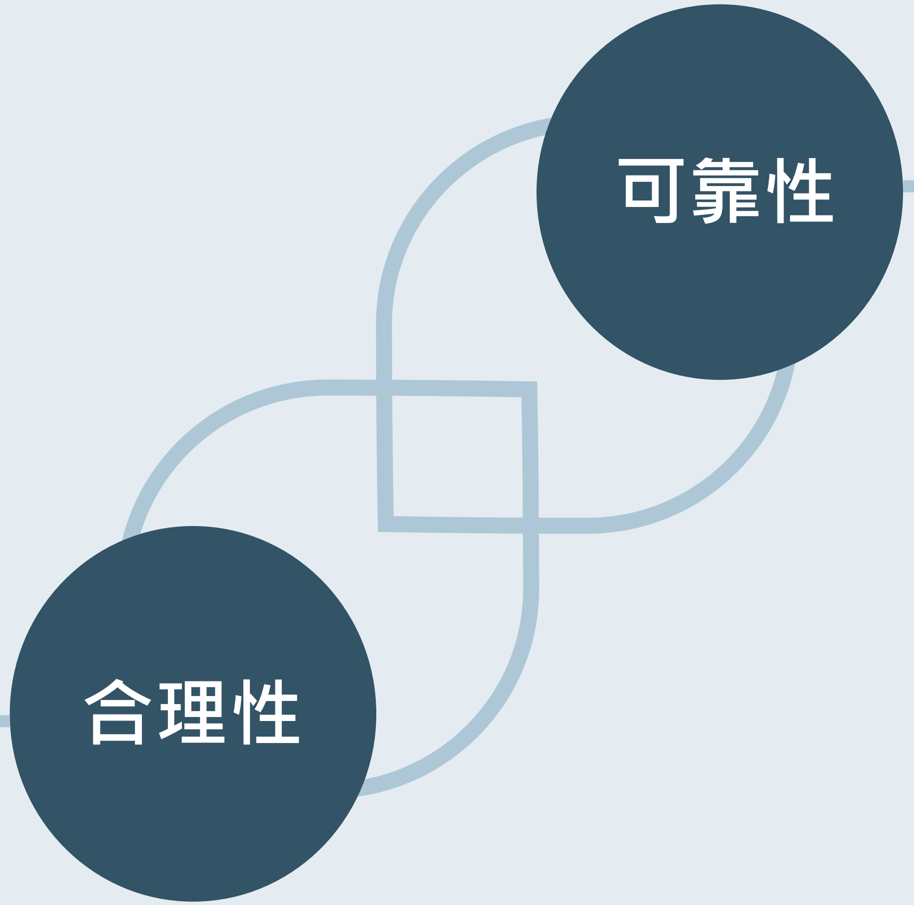
- 為什麼不使用**詞向量模型 (Word2Vec)**？

  ➔ 因為詞向量模型是以生成正確答案為目標，並不適合成為干擾選項。

干擾選項的標準

Plausibility

干擾選項和答案必須**語意上相近、文法上相同**，使作答者容易被干擾選項混淆，增加題目鑑別度



Reliability

干擾選項**不能和答案意思相同**，避免干擾選項也可以當作答案

干擾選項的標準

可靠性

困難點：之前的模型無法同時滿足這兩個標準

合理性

相關研究

透過語料庫(如Probase[6], Wordnet[7])生成干擾選項

Problem：語料庫不存在該單字的問題

先蒐集特定領域的相關詞彙並建立字典

Problem：特定領域的字典需要人工收集與建立，必須付出較高成本

相關研究

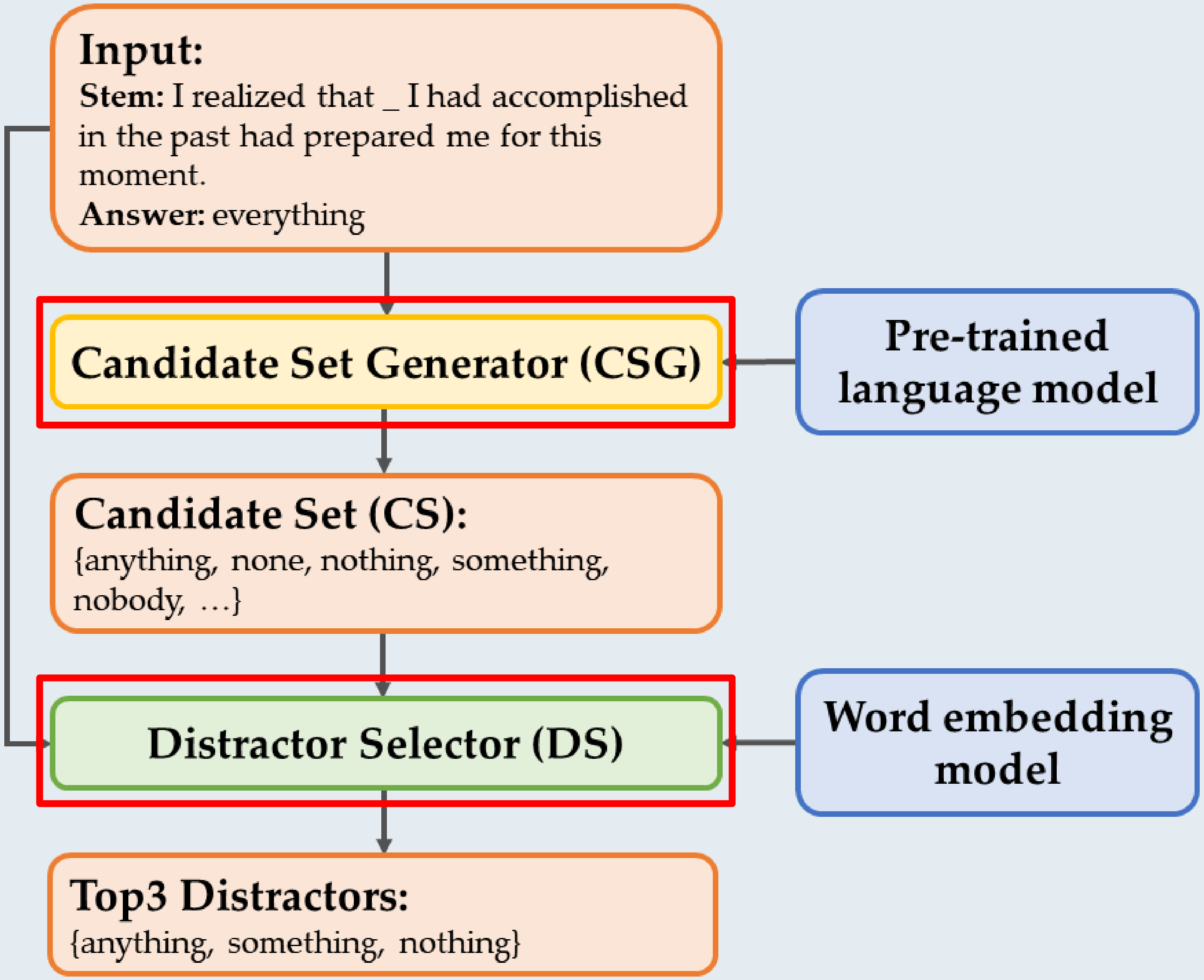
目標

研究一種可以**跨領域**且**自動化**的方法，在符合**可靠性**和**合理性**的前提下，滿足克漏字干擾選項生成的實際應用。

克漏字資料集與自動評估指標

- 實驗中我們主要引用了**CLOTH**[11]資料集進行訓練。
- 計算Precision，Recall，F1 score、MRR以及NDCG@10來衡量表現。

CDGP框架



Input:
Stem: I realized that _ I had accomplished in the past had prepared me for this moment.
Answer: everything
Candidate Set Generator (CSG)
Pre-trained language model
Candidate Set (CS):
{anything, none, nothing, something, nobody, …}
Distractor Selector (DS)
Word embedding model
Top3 Distractors:
{anything, something, nothing}

1. 前言
2. Candidate Set Generator (CSG)
3. Distractor Selector (DS)
4. 相關實驗
5. 結論
6. 參考文獻
7. Q&A

Candidate Set Generator (CSG)

考題和答案 ⇨ CSG ⇨ Candidate Set (CS)：10個最合適的干擾選項

Candidate Set Generator (CSG)

CSG

經過微調的語言預訓練模型

- 我們使用**Mask-filling**方式進行微調
- 根據微調的方法不同，又分成兩種：
  1. Normal
  2. Answer Relating

CSG訓練方法

- Normal

考題：__ , Jane didn't understand her. 答案：However 干擾選項：Though、Although、Or
輸入：[MASK], Jane didn't understand her. 標籤1：Though, Jane didn't understand her. 標籤2：Although, Jane didn't understand her. 標籤3：Or, Jane didn't understand her.

- 希望模型可以從訓練中學會**預測挖空部分適合的干擾選項**。

CSG訓練方法

- Answer Relating

考題：__ , Jane didn't understand her.
答案：However
干擾選項：Though、Although、Or
輸入：[MASK], Jane didn't understand her. [SEP] However
標籤1：Though, Jane didn't understand her. [SEP] However
標籤2：Although, Jane didn't understand her. [SEP] However
標籤3：Or, Jane didn't understand her. [SEP] However

- 希望模型可以從訓練中學會找出人類出題時，**干擾選項和答案間的關聯**。

Normal和Answer Relating之比較

Models	P@1	F1@3	F1@10	MRR	NDCG @10
Normal	12.60	10.00	12.45	22.70	30.32
Answer Relating	18.50	13.80	15.37	29.96	37.82

- Answer Relating 的結果優於 Normal，之後皆延續此方法。

1. 前言
2. Candidate Set Generator (CSG)
3. Distractor Selector (DS)
4. 相關實驗
5. 結論
6. 參考文獻
7. Q&A

Distractor Selector (DS)

考題、答案
和CS

⇨

DS

⇨

Top3 Distractors：
取分數最高的前三名
作為最終干擾選項

Distractor Selector (DS)

DS

關於排名的分數，我們挑選出以下四個評分指標：

- **S0(模型信心分數)**：CSG生成結果，信心值越高，分數越高。
- **S1(單字相似度)**：和答案的單字意思越接近，分數越低。
- **S2(句子相似度)**：和答案的句子意思越接近，分數越低。
- **S3(詞性相似度)**：和答案詞性相同，分數越高。

1. 前言
2. Candidate Set Generator (CSG)
3. Distractor Selector (DS)
4. 相關實驗
5. 結論
6. 參考文獻
7. Q&A

框架使用與否之比較

- CDGP實際上到底有沒有提升模型表現呢？

Methods	P@1	F1@3	F1@10	MRR	NDCG @10
CSG+DS	**19.30**	**15.50**	**15.37**	**31.26**	**39.49**
CSG	18.50	14.90	15.37	30.57	38.73
DS	4.00	6.43	5.05	12.02	19.12
None	4.10	6.03	5.05	11.81	18.65

- 相比於沒有使用的模型，**效果大幅提升**，證明CDGP在干擾選項生成有著卓越的效果。

與其他研究之比較

- 我們參考克漏字干擾選項生成最新的研究[1]，並且找到此研究所使用的資料集 - DGen。
- 該資料集包含許多科學相關文章與問題，所以預訓練模型改用經過許多科學文章訓練過的**SciBERT**模型[23]。

Models	P@1	F1@3	MRR	NDCG@10
研究[1]之最好數據	10.85	9.19	17.51	19.31
SciBERT_DGen	13.13	12.23	25.12	34.17

- 我們的**NDCG@10**從**19.31**提升至**34.17**，**超越現有方法77%**，可見表現上已明顯超越目前最新的研究方式。

人工評估

- 40位志願者
- 讀一篇英文文章並完成10題克漏字，5題為人工出題，5題CDGP出題

人工評估

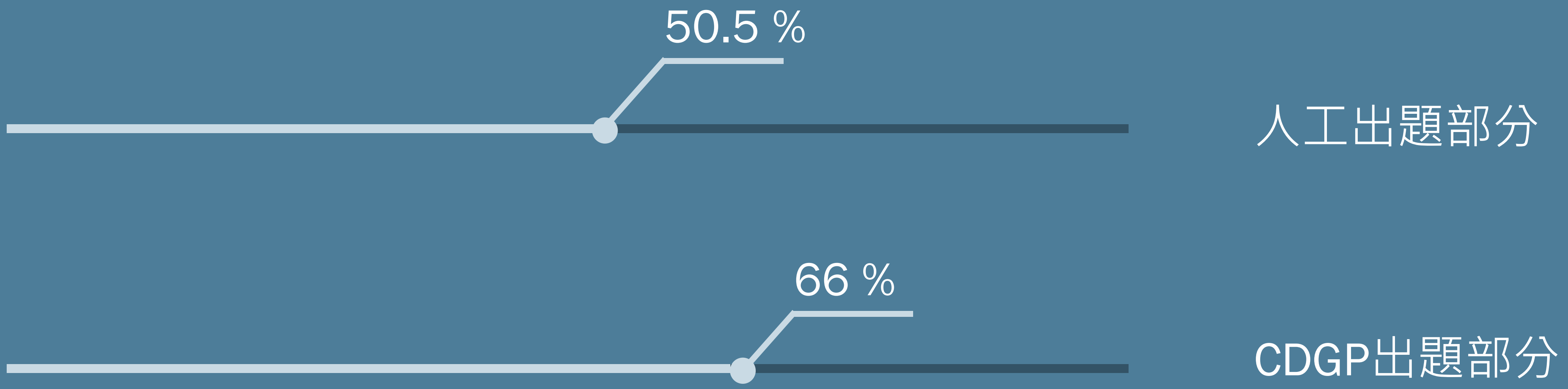
- 答題正確率：

人工出題部分

CDGP出題部分

人工評估

- 答題正確率：



人工評估

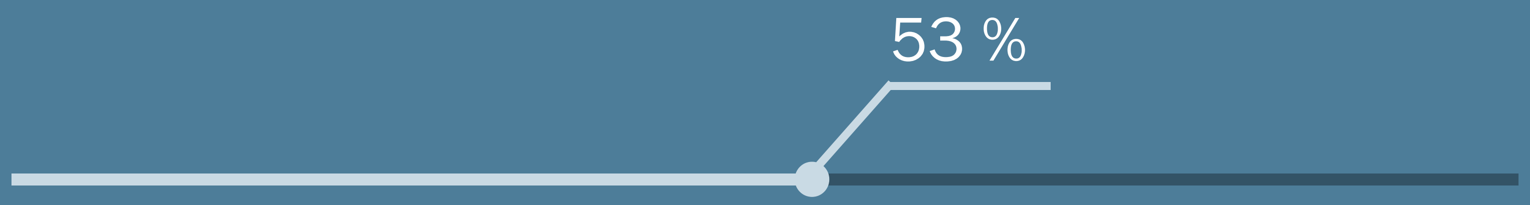
- 分辨人工與CDGP出題方面

從10題中猜5題為CDGP

出題的猜中率

人工評估

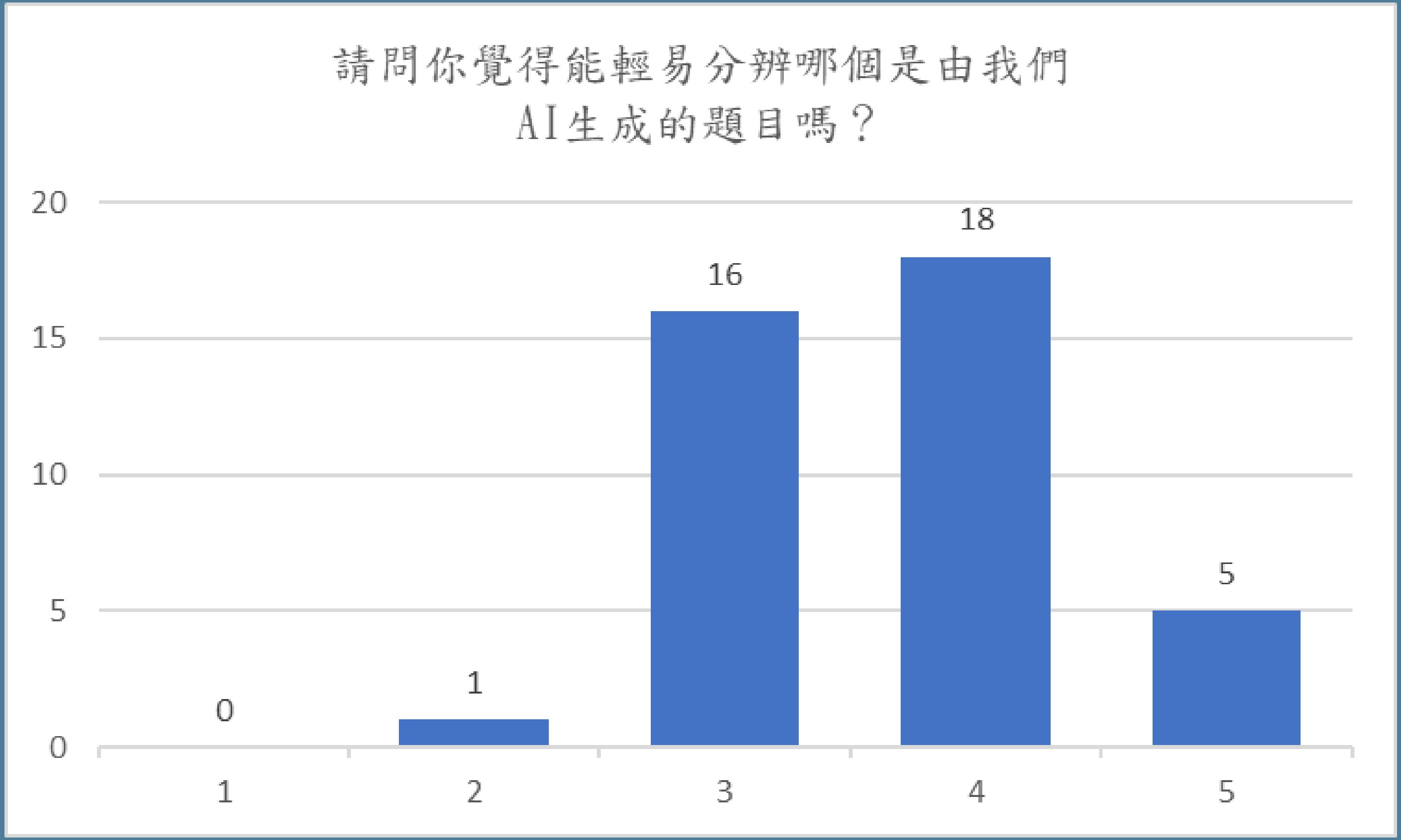
- 分辨人工與CDGP出題方面



從10題中猜5題為CDGP出題的猜中率

人工評估

- 測試者對是否能分辨人工與CDGP出題差別的感想 (1:容易分辨，5:難以分辨)



人工評估

- CDGP出題的表現非常接近**人工出題**，證實CDGP在克漏字干擾選項的生成是有能力去輔佐人類的。

考題	If it is more money that you want, start being grateful for whatever __ of money you already have.		
答案	**amount**		
人工生成	**kind** **number** plenty	CDGP生成	**kind** **number** type

1. 前言
2. Candidate Set Generator (CSG)
3. Distractor Selector (DS)
4. 相關實驗
5. 結論
6. 參考文獻
7. Q&A

結論

- 證實基於**語言預訓練模型**，經過微調後在干擾選項生成方面，相比於使用語料庫(Probase[6], Wordnet[7])，能有更好的效果。
- 進一步提出**Answer Relating**的方法，藉由**學習干擾選項與答案間的關聯**，來提升模型生成的效果。
- 在與研究[1]的比較中，**CDGP大幅度地勝過現有最好之生成效果，將NDCG@10 提升了77%**。

Cloze Distractor Generator

Please enter an article for cloze test:

Confirm ✓

Customize

Stars 0


1
前言
2
Candidate Set Generator (CSG)
3
Distractor Selector (DS)
4
相關實驗
5
結論
6
參考文獻
7
Q&A

參考文獻

[1] Ren, S., & Q. Zhu, K. (2021). Knowledge-Driven Distractor Generation for Cloze-Style Multiple Choice Questions. Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4339-4347
[2] KALPAKCHI, Dmytro; BOYE, Johan. BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. *arXiv preprint arXiv:2108.03973*, 2021.
[3] OFFERIJNS, Jeroen; VERBERNE, Suzan; VERHOEF, Tessa. Better distractions: Transformer-based distractor generation and multiple choice question filtering. arXiv preprint arXiv:2010.09598, 2020.
[4] CHUNG, Ho-Lam; CHAN, Ying-Hong; FAN, Yao-Chung. A BERT-based Distractor Generation Scheme with Multi-tasking and Negative Answer Training Strategies. In: Findings of the Association for Computational Linguistics: EMNLP 2020. 2020. p. 4390-4400.
[5] CHAN, Ying-Hong; FAN, Yao-Chung. A recurrent BERT-based model for question generation. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. 2019. p. 154-162.
https://classic.queratorai.com/
[6] WU, Wentao, et al. Probase: A probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. 2012. p. 481-492.
[7] MILLER, George A. WordNet: a lexical database for English. Communications of the ACM, 1995, 38.11: 39-41.
[8] KUMAR, Girish; BANCHS, Rafael E.; D’ HARO, Luis Fernando. Revup: Automatic gap-fill question generation from educational texts. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. 2015. p. 154-161.
[9] JIANG, Shu; LEE, John SY. Distractor generation for chinese fill-in-the-blank items. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. 2017. p. 143-148.
[10] LAI, Guokun, et al. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683, 2017.
[11] XIE, Qizhe, et al. Large-scale cloze test dataset created by teachers. arXiv preprint arXiv:1711.03225, 2017.

Workshop on Innovative Use of NLP for Building Educational Applications. 2017. p. 143-148.
[10] LAI, Guokun, et al. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683, 2017.
[11] XIE, Qizhe, et al. Large-scale cloze test dataset created by teachers. arXiv preprint arXiv:1711.03225, 2017.
[12] LOPER, Edward; BIRD, Steven. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.
[13] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[14] LEWIS, Mike, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.
[15] LIU, Yinhan, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
[16] WELBL, Johannes; LIU, Nelson F.; GARDNER, Matt. Crowdsourcing multiple choice science questions. arXiv preprint arXiv:1707.06209, 2017.
https://allenai.org/data/sciq
[17] LIANG, Chen, et al. Distractor generation for multiple choice questions using learning to rank. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications. 2018. p. 284-290.
[18] AI2 Science Questions
http://data.allenai.org/ai2-science-questions/
[19] Hugging Face, bert-base-uncased. https://huggingface.co/bert-base-uncased
[20] WU, Sun; MANBER, Udi. Fast text searching: allowing errors. *Communications of the ACM*, 1992, 35.10: 83-91.
[21] Hugging Face, facebook/bart-base. https://huggingface.co/facebook/bart-base
[22] Hugging Face, roberta-base.
https://huggingface.co/roberta-base
[23] Hugging Face, allenai/scibert_scivocab_uncased. https://huggingface.co/allenai/scibert_scivocab_uncased
[24] HSU, Tsung-Yuan; LIU, Chi-Liang; LEE, Hung-yi. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. *arXiv preprint arXiv:1909.09587*, 2019.

1. 前言
2. 候選字集合生成器 (CSG)
3. 干擾選項生成器 (DS)
4. 實驗與結論
5. 研究貢獻
6. 參考文獻
7. Q&A

Q&A

Thanks for watching!