Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 67-84

Answering Chinese Elementary School Social Studies Multiple Choice Questions

Chao-Chun Liang[*], Daniel Lee[†], Meng-Tse Wu[‡],

Hsin-Min Wang[*], and Keh-Yih Su[*]

Abstract

We present several novel approaches to answer Chinese elementary school social studies multiple choice questions. Although BERT shows excellent performance on various reading comprehension tasks, it handles some kinds of questions poorly, in particular *negation*, *all-of-the-above*, and *none-of-the-above* questions. We thus propose a novel framework to cascade BERT with preprocessor and answer-picker/selector modules to address these cases. Experimental results show the proposed approaches effectively improve the performance of BERT, and thus demonstrate the feasibility of supplementing BERT with additional modules.

Keywords: Natural Language Inference, Machine Reading Comprehension, Multiple Choice Question, Question and Answering.

1. Introduction

Machine reading comprehension (MRC) is a challenge for AI research, and is frequently adopted to seek desired information from knowledge sources such as company document collections, Wikipedia or the Web for a given question. To evaluate the capability of a MRC system, different test forms have been adopted in the literature (Qiu *et al.*, 2019; Liu *et al.*, 2019) such as binary choice, multiple choice (MC), multiple selection (MS), and cloze. Which test form to adopt usually depends on the format of the given benchmark/dataset. In this paper,

[*] Institute of Information science, Academia Sinica, Taipei, Taiwan
E-mail: {ccliang, whm, kysu}@iis.sinica.edu.tw

[†] Department of Computer Science Engineering, University of Michigan, Ann Arbor, Michigan, USA (His work was done during his summer internship in Institute of Information science, Academia Sinica).
E-mail: danclee@umich.edu

[‡] NYU Tandon School of Engineering, Brooklyn, NY, USA (His work was done during his research assistantship in Institute of Information science, Academia Sinica).
E-mail: michaelmoju@gmail.com

Table 1. Example social studies MC question.

Passage	三代同堂家庭是子女和父母、祖父母或外祖父母同住。
Question	「我和爸爸、媽媽、爺爺、奶奶住在一起。」是屬於哪一種類型的家庭？
Options	(1) 三代同堂家庭 (2) 單親家庭 (3) 隔代教養家庭 (4) 寄養家庭
Answer	(1) 三代同堂家庭

we solve MC questions about traditional Chinese primary school social studies. In this Chinese Social Studies MC (CSSMC) QA task, the system selects the correct answer from several candidate options based on a given question and its associated lesson manually constructed by Taiwan book publishers. Table 1 shows an example of CSSMC, where the passage is the corresponding supporting evidence (SE).

Previous work on answering MC questions can be divided into statistics-based approaches (Kouylekov & Magnini, 2005; Heilman & Smith, 2010) and neural-network-based approaches (Parikh *et al*., 2016; Chen *et al*., 2017). Recent pre-trained language models such as BERT (Devlin *et al*., 2019), XLNET (Yang *et al*., 2019), RoBERTa (Liu *et al*., 2019), and ALBERT (Lan *et al*., 2019) show excellent performance on different RC MC tasks. As BERT shows excellent performance on various English datasets (e.g., SQuAD 1.1 (Rajpurkar *et al*., 2016), GLUE (Wang *et al*., 2018), etc.), it is adopted as our baseline. Table 6 shows its performance given the gold SE.

After analyzing error cases, we observed that BERT handles the following question types poorly: (1) **Negation** questions, that is, questions with negation phrases such as 不可能 (unlikely). For this type of question, BERT selects the same answer for “小敏的媽媽目前在郵局服務，請問小敏的媽媽**可能**會為居民提供什麼服務？ (Xiaomin's mother serves at the post office. What kind of services could Xiaomin's mother provide to the residents?)” and “小敏的媽媽目前在郵局服務，請問小敏的媽媽**不可能**會為居民提供什麼服務？ (Xiaomin's mother serves at the post office. What kind of service could **not** Xiaomin's mother provide to the residents?)” (which differ only in the negation word 不 (**not**)). BERT evidently pays no special attention to negative words; however, any one of them would change the desired answer; (2) **All-of-the-above** (以上皆是) and **none-of-the-above** (以上皆非) questions, choices for which include either *All of the above* or *None of the above*. In both cases, the answer cannot be handled by simply by selecting the most likely choice without preprocessing

Table 2. Question types in CSSMC corpus.

Problem type	Questions
Negation	**Question**: 浩浩跟家人到臺東縣關山鎮遊玩，他不可能在當地看到什麼？ **Options**: (1)阿美族豐年祭 (2)環鎮自行車道 (3)油桐花婚禮 (4)親水公園
All of the above	**Question**: 在高齡化的社會裡，我們應該如何因應高齡化社會的到來？ **Options**: (1)制定老人福利政策 (2)提供良好的安養照顧 (3)建立健全的醫療體系 (4)以上皆是
None of the above	**Question**: 都市有公共設施完善、工作機會多等優點，常吸引鄉村地區哪一種年齡層的居民前往？ **Options**: (1)老人 (2)小孩 (3)青壯年 (4)以上皆非

the given choices. Table 2 shows an example of these question types.

The above phenomenon was also observed by Wu & Su (2020), who reported that BERT achieves superior results mainly by utilizing surface features, and that its performance degrades significantly when the dataset involves negation words. Moreover, it is difficult for BERT to learn the semantic meaning of all-of-the-above and none-of-the-above questions, which suggests that the listed candidate options are all correct or all incorrect, with a small amount of data.

However, it is difficult to pinpoint the sources of the problem and then find corresponding remedies within BERT, due to its complicated architecture (even its basic version includes 12 heads and 12 stacked layers). We thus prefer to keep its implementation untouched if the problem can be fixed by coupling BERT with external modules. Accordingly, we here propose a framework that cascades BERT with a preprocessor module and an answer-picker/selector module. The preprocessor module revises the choices for all-of-the-above and none-of-the-above questions, and the answer-picker/selector module (a postprocessor) determines the appropriate choices under the cases mentioned above. The above approach is inspired by Lin & Su (2021), who demonstrate that BERT learns natural language inference inefficiently, even for simple binary prediction; however, they also point out that task-related features and domain knowledge significantly help to improve BERT's learning efficiency.

For negation-type questions, instead of picking the highest-scoring choice as usual, the answer-picker/selector module selects the candidate with the lowest score. On the other hand, for all-of-the-above or none-of-the-above questions, we use a decision tree to select the answer, as illustrated in Figure 2. In these cases, the preprocessor module first replaces the original “all of the above” or “none of the above” choices with a new choice generated by concatenating all other choices together (before those candidates are sent to BERT). Take for example the second last row in Table 2: we replace “以上皆是 (all of the above)”, the original last choice, with “制定老人福利政策^提供良好的安養照顧^建立健全的醫療體系 (Make welfare policies for elderly people^ Provide good nursing care^ Establish a sound medical system)”.

We evaluate the proposed framework on a CSSMC dataset. The experimental results show the proposed approaches outperform the pure BERT model. This thus constitutes a new way to supplement BERT with additional modules. We believe the same strategy could be applied to other DNN models, which — despite good overall performance — are too complicated to customize for specific problems.

In summary, in this paper we make the following contributions: (1) We propose several novel approaches to supplement BERT to solve negation, all-of-the-above, and none-of-the-above questions. (2) Experimental results show that the proposed approach effectively improves performance, and thus demonstrate the feasibility of supplementing BERT with additional modules to fix given problems. (3) We construct and release a new Traditional Chinese Machine Reading Question and Answering dataset to assess the performance of RC MC models.

In comparison with our previous conference version (Lee *et al*., 2020), this article describes additional “*Separately Judge then Select*” and “*Separately Judge Concatenation then Select*” experiments, which adopt a BERT entailment prediction model to handle each candidate option separately (details are provided in Sections 2.2.1 and 2.2.2) instead of jointly processing all candidate options together. We have also added Section 3 to describe the construction of the CSSMC dataset, which we adopt to compare different approaches.

2. Proposed Approaches

2.1 Problem Formulation

Given a social studies problem *Q* and its corresponding supporting evidence *SE*, our goal is to find the most likely answer from the given candidate set $A = \{A_1, A_2, \dots A_n\}$, where *n* is the total number of available choices or candidates, and A_i denotes the *i*-th answer candidate. This task is formulated as follows, where $\hat{A}$ is the answer to be chosen.

$$\hat{A} = \underset{i=1,\dots,n}{\arg\max}\, P(\mathrm{A}_i \mid Q, SE, A) \tag{1}$$

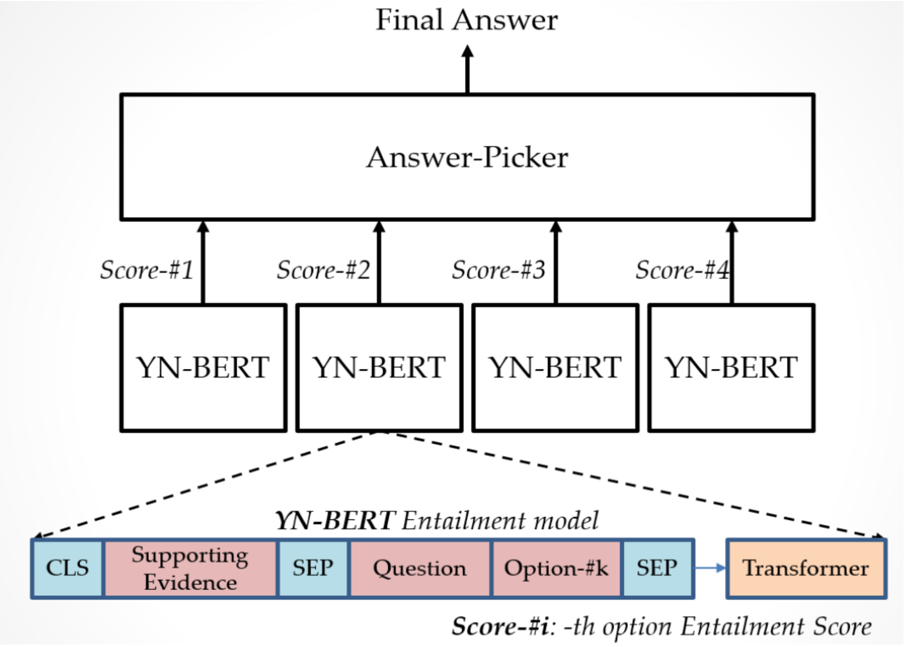


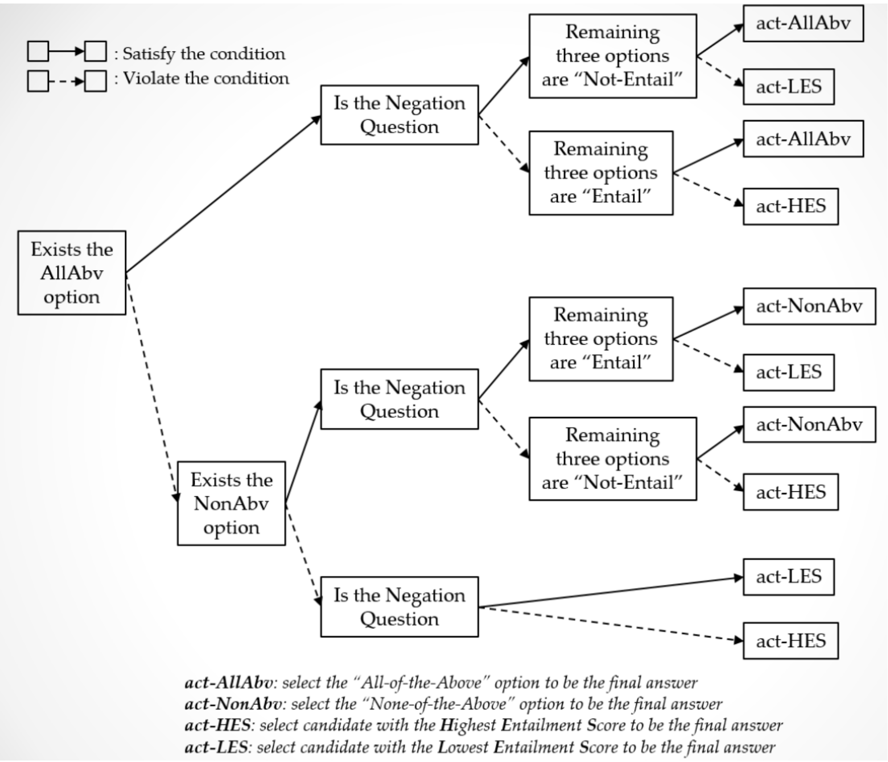
Figure 1. Architecture of proposed SJS approach.

2.2 Proposed Models

Three different approaches are proposed in which we use entailment prediction (Dagan *et al.*, 2005) to determine whether the candidate option is the correct answer to the question: (1) *Separately judge then select* (SJS), which considers each individual candidate option separately and then selects the final answer based on their output scores; (2) *Separately judge with concatenation then select* (SJCS), which adopts the framework of the first approach but first replaces the all-of-the-above (以上皆是) and none-of-the-above (以上皆非) answer choices with the concatenation of all the other remaining candidate options before entailment judgment; (3) *Jointly judge then select* (JJS), which jointly considers all candidate options to make the final decision. Details are provided below.

2.2.1 Separately Judge then Select (SJS)

Figure 1 shows the architecture of the proposed SJS approach, which consists of two main components: (1) the YN-BERT module, a fine-tuned BERT entailment prediction model (where YN denotes its output is a yes-no binary entailment judgment), and (2) the answer-picker module, which determines the final answer given the entailment judgment scores from four different YN-BERT modules. The input sequence is the concatenation of the associated supporting evidence, a given question, and a specific individual answer candidate/option. For each answer candidate, YN-BERT outputs an entailment judgment score

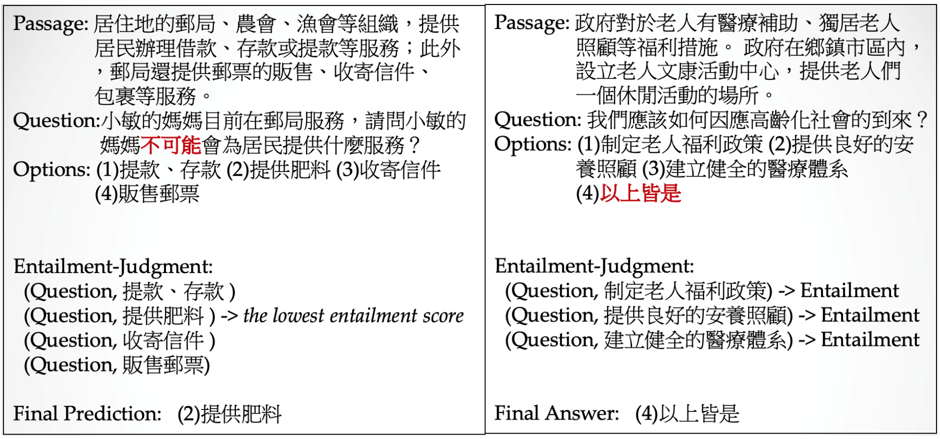


: Satisfy the condition
: Violate the condition
Exists the AllAbv option
Is the Negation Question
Remaining three options are "Not-Entail"
act-AllAbv
act-LES
Remaining three options are "Entail"
act-AllAbv
act-HES
Exists the NonAbv option
Is the Negation Question
Remaining three options are "Entail"
act-NonAbv
act-LES
Remaining three options are "Not-Entail"
act-NonAbv
act-HES
Is the Negation Question
act-LES
act-HES
act-AllAbv: select the "All-of-the-Above" option to be the final answer
act-NonAbv: select the "None-of-the-Above" option to be the final answer
act-HES: select candidate with the Highest Entailment Score to be the final answer
act-LES: select candidate with the Lowest Entailment Score to be the final answer

Figure 2. Decision tree for SJS approach. Each "act-xxx" is a specific action to be taken.

used to select either *Entail* or *Not-entail* (i.e., the judgment is *Entail* if the score exceeds 0.5, and *Not-entail* otherwise). *Entail* implies that the given answer candidate is entailed by the combination of the question and its associated supporting evidence. The answer-picker module considers the entailment judgment scores of the various choices and selects the most appropriate one based on the decision tree shown in Figure 2. Note that this decision tree is used only by the answer picker to make the final decision and is not involved in BERT's fine-tuning process.

A given question is classified as negative-type if it includes a negation word within a pre-specified negation word list, which is obtained from the CSSMC training data, and currently consists of {"不會 (will not)", "不能 (cannot)", "不得 (not allow)", "不是 (is not)", "不應該 (should not)", "不可能 (unlikely)", "不需 (do not need)", "不必 (do not need)", "不用 (do not need)", "沒有 (without)"}. Since the proposed approaches aim to supplement BERT, these negation words are manually picked from the error cases in the training data-set, on which BERT model make mistakes. Figure 3 shows the examples under two different inference mechanisms: (1) for a negation-type question (left figure), and (2) a question with all of the above option (right figure).



Passage: 居住地的郵局、農會、漁會等組織，提供居民辦理借款、存款或提款等服務；此外，郵局還提供郵票的販售、收寄信件、包裹等服務。
Question:小敏的媽媽目前在郵局服務，請問小敏的媽媽不可能會為居民提供什麼服務？
Options: (1)提款、存款 (2)提供肥料 (3)收寄信件 (4)販售郵票

Entailment-Judgment:
(Question, 提款、存款)
(Question, 提供肥料) -> *the lowest entailment score*
(Question, 收寄信件)
(Question, 販售郵票)

Final Prediction: (2)提供肥料

Passage: 政府對於老人有醫療補助、獨居老人照顧等福利措施。 政府在鄉鎮市區內，設立老人文康活動中心，提供老人們一個休閒活動的場所。
Question: 我們應該如何因應高齡化社會的到來？
Options: (1)制定老人福利政策 (2)提供良好的安養照顧 (3)建立健全的醫療體系 (4)**以上皆是**

Entailment-Judgment:
(Question, 制定老人福利政策) -> Entailment
(Question, 提供良好的安養照顧) -> Entailment
(Question, 建立健全的醫療體系) -> Entailment

Final Answer: (4)以上皆是

Figure 3. Two inference mechanisms under SJS framework.

2.2.2 Separately Judge with Concatenation then Select (SJCS)

Another approach adopts the framework of the first approach but first recasts “以上皆是 (all of the above)” and “以上皆非 (none of the above)’ answer candidates as the concatenation of all of the other options. Take for example the last row in Table 2: we replace “以上皆非”, the original last choice, with “老人^小孩^青壯年 (elderly people^children^young people)”. Afterwards, the answer-picker module selects the most appropriate choice based on the following rule: For negation questions, we select the answer candidate with the lowest entailment score; otherwise, we select that with the highest entailment score.

2.2.3 Jointly Judge then Select (JJS)

Shown in Figure 4, the system architecture of the JJS approach consists of three main components: (1) the preprocessor, which recasts “以上皆是 (all of the above)” and “以上皆非 (none of the above)” answer candidates as the concatenation of the other options (associated with the same question), as shown above, before inputting the question-choice-evidence combination into the BERT model; (2) the BERT-MC model, a typical fine-tuned BERT multiple-choice prediction model (Xu *et al*., 2019) described in Section 4.1; and (3) the answer selector, a candidate re-selector which for negation-type questions picks that answer candidate with the lowest score as opposed to that with the highest score (as for other question types).

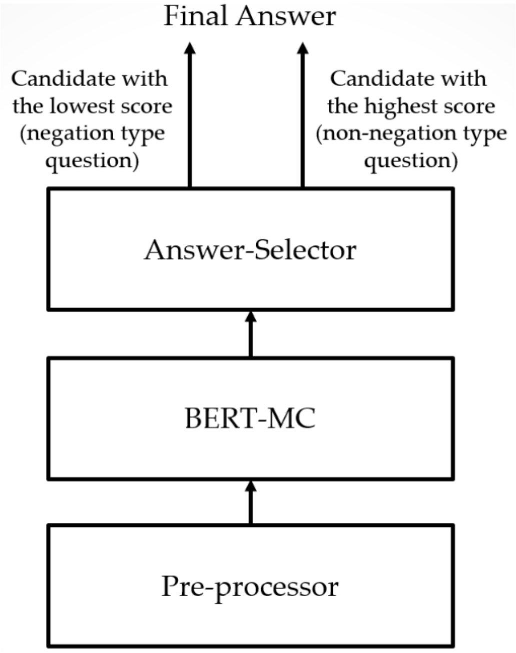

Final Answer
Candidate with the lowest score (negation type question)
Candidate with the highest score (non-negation type question)
Answer-Selector
BERT-MC
Pre-processor

Figure 4. System architecture of proposed "Jointly Judge then Select" framework.

3. Chinese Social Studies MRQA Dataset Construction

To evaluate the proposed approaches, we constructed a *Chinese Social Studies Machine Reading and Question Answering* (CSSMRQA) dataset, which is a superset of the CSSMC dataset mentioned above, to assess the capability of different Q&A systems (not just MC questions). This dataset consists of three question types: (1) yes/no questions, which ask whether the given question is a correct statement judged from the supporting evidence; (2) multiple-choice (MC) questions, which include four answer choices from which the correct one is to be chosen (here, this is the CSSMC dataset adopted in this paper); and (3) multiple-selection (MS) questions, which are similar to the multiple-choice questions but can contain more than one correct answer. Below we describe how they are constructed.

3.1 Corpus Collection

We first collected lessons for grades 3 to 6 from elementary-school social studies textbooks published in Taiwan. For each lesson, we collected relevant questions from leading publishing houses in Taiwan. We thus obtained 14,103 yes/no questions, 5347 MC questions, and 340 MS questions from a total of 255 lessons. We then annotated the supporting evidence to indicate what information is needed to answer each question. This is described in detail below.

3.2 Supporting Evidence (SE) Annotation

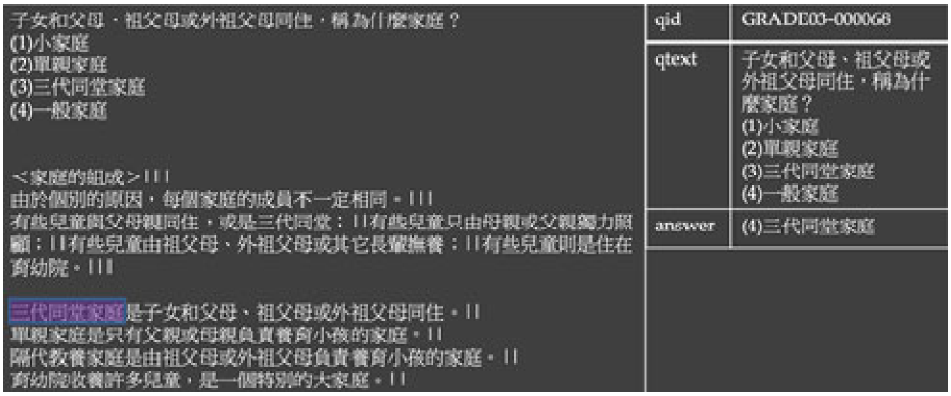
We hired two annotators to annotate the supporting evidence for each question. Supporting evidence is the content in the lesson (associated with the given question) which contains just the information necessary to answer the question. In the CSSMRQA dataset, each lesson comprises several paragraphs, and each paragraph comprises several sentences. Supporting evidence consists of one or more sentences.

We used Doccano (Nakayama *et al.*, 2018), an open-source text annotation tool, as the platform for annotation. Doccano allows the user to highlight supporting words in the text (i.e., those words that provide hints to find the related passage). Given a question and its corresponding answer (also the lesson associated with the question), the annotators highlighted supporting words necessary to answer the question. Usually, these supporting words were words within the given question. Annotators were not allowed to annotate supporting words across sentence splitters or delimiters. Nonetheless, some questions lack suitable supporting evidence in the lesson. For example, students may rely on common sense (instead of textbook context) to answer the question, "班上同學有人亂丟垃圾，身為衛生股長的小玉可以怎麼做？ (1) 默默的跟在他們後面撿垃圾 (2) 勸告亂丟垃圾的同學，並請他們將垃圾撿起來 (3) 沒關係，等打掃時間再掃就好了 (4) 把垃圾藏在看不見的地方 (What can Xiaoyu (the Chief of Health) do when her classmate litters? (1) Pick up trash after them silently; (2) Advise the classmate who litters and ask him/her to pick up the litter; (3) It doesn't matter, just wait until the cleaning time; or (4) Hide litter out of sight)". In such cases, annotators found no suitable supporting words in the lesson and thus skipped SE annotation. Afterward, sentences that contain marked supporting words were annotated as supporting evidence. Table 3 shows the final results of SE annotation.

Table 3. Supporting evidence annotation in training/dev/test subsets.

Subset	Training	Dev	Test
Questions	3,879	780	778
Questions w/ SE	3,135	604	563
Questions w/o SE	744	176	215
Averaged SPs	1.09	1.16	1.14
Averaged SSs	3.17	2.94	2.73

*Questions w/o SE: the number of questions without supporting evidence
Averaged SPs: the average number of Supporting Paragraphs
Averaged SSs: the average number of Supporting Sentences



子女和父母、祖父母或外祖父母同住，稱為什麼家庭？
(1)小家庭
(2)單親家庭
(3)三代同堂家庭
(4)一般家庭

＜家庭的組成＞|||
由於個別的原因，每個家庭的成員不一定相同。|||
有些兒童與父母親同住，或是三代同堂；||有些兒童只由母親或父親獨力照顧；||有些兒童由祖父母、外祖父母或其它長輩撫養；||有些兒童則是住在育幼院。|||

三代同堂家庭是子女和父母、祖父母或外祖父母同住。||
單親家庭是只有父親或母親負責養育小孩的家庭。||
隔代教養家庭是由祖父母或外祖父母負責養育小孩的家庭。||
育幼院收養許多兒童，是一個特別的大家庭。||

qid	GRADE03-000068
qtext	子女和父母、祖父母或外祖父母同住，稱為什麼家庭？ (1)小家庭 (2)單親家庭 (3)三代同堂家庭 (4)一般家庭
answer	(4)三代同堂家庭

Figure 5. Multiple-choice question annotation.

Figure 5 shows an example of multiple-choice question annotation. Annotators first read both the question (qtext) and the correct answer (answer) from the right-hand side windows, and then highlight supporting words (marked with purple boxes) in the lesson. To prevent annotators from highlighting supporting word regions across sentences, we use special symbols as separators (||| for paragraphs and || for sentences).

4. Experiments

We conducted experiments on the above CSSMC dataset with the three proposed approaches. Table 4 shows the dataset statistics. For comparison, we used a typical BERT multiple-choice implementation (Xu *et al*., 2019) as our baseline.

4.1 Baseline: BERT-MC

For the baseline, we used the BERT-MC model, that is, BERT (Devlin *et al*., 2019) fine-tuned for the multiple-choice task as our baseline, as it is the most widely adopted state-of-the-art model (Xu *et al*., 2019). It was built by exporting BERT's final hidden layer into a linear layer and then taking a softmax operation. For details on the BERT-MC model, please see Xu *et al*. (2019). The BERT input sequence consists of "[CLS] SE [SEP] Question [SEP] Option-#*i* [SEP]", where Option-#*i* denotes the *i*-th option and [CLS] and [SEP] are special tokens representing the classification and the passage separators, respectively, as defined in Devlin *et al*. (2019). Figure 6 shows the architecture of the BERT baseline model.

Table 4. CSSMC dataset.

	Training	Dev	Test
Lessons	202	27	26
Questions	3,879	780	778
Averaged paragraphs/lesson	11.28	13.93	10.93
#Averaged entences/lesson	46.40	52.67	46.33

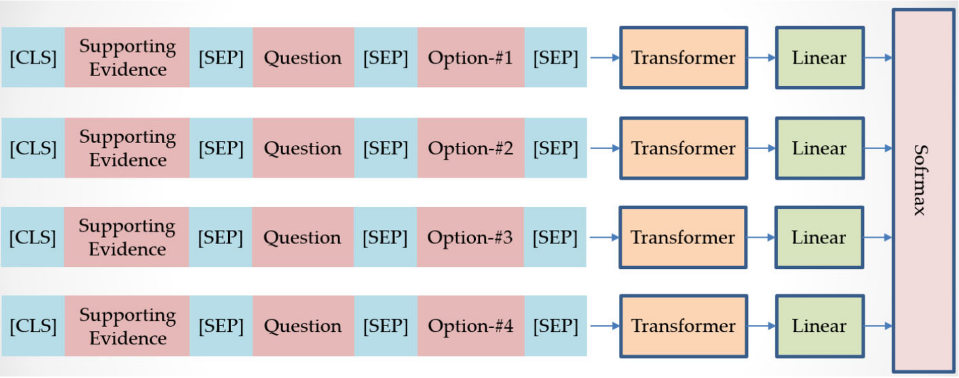


Figure 6. The architecture of the BERT-MC model (Xu et al., 2019).

4.2 Retrieved Supporting Evidence (SE) Dataset

SE is the corresponding shortest passage based on which the system can answer the given question. Given the annotation results described in Section 3.2, we find many questions that involve common-sense reasoning, for which no corresponding SEs can be found in the retrieved lesson. We denote as SE1 that set of questions for which SEs can be found in the retrieved lesson (this is termed GSE1 if it is also associated with gold SEs); the set of remaining questions is SE2. Table 5 shows the statistics for GSE1.

Table 5. CSSMS GSE1 (with gold SEs) subset statistics.

	Training	Dev	Test
Lessons	196	27	26
Questions (NEG[a]) (AllAbv&NonAbv[b])	3,135 (53) (332)	604 (14) (69)	563 (15) (56)
Averaged paragraphs/lesson	11.35	13.93	10.85
Averaged sentences/ Lesson	46.72	52.67	46.15

[a] NEG: number of negation-type questions.
[b.] AllAbv&NonAbv: number of AllAbv&NonAbv-type questions.

4.3 Results

We conducted two sets of experiments on the CSSMC dataset: (i) GSE1, based on SE1 with gold SEs, to compare the QA component performance of different models; and (ii) LSE, based on the whole dataset with all SEs directly retrieved from the Lucene search engine, to compare different approaches under a real-world situation. Each set covers six different models: (1) *BERT-MC Only*, (2) *SJS*, (3) *SJCS*, (4) *BERT-MC+Neg*, (5) *BERT-MC+AllAbv&NonAbv*, and (6) *BERT-MC+Neg+AllAbv&NonAbv*, where *BERT-MC Only* is the baseline model and *Neg* and *AllAbv&NonAbv* denote additional answer-selector and preprocessor modules for the negation and all-of-the-above/none-of-the-above question-types, respectively. We adopted the setting specified in Xu *et al.* (2019) for BERT training. All other models were trained using the following hyperparameters: (1) a maximum sequence length of 300; (2) a learning rate of 5e-5 with the AdamW optimizer (Loshchilov & Hutter, 2019); (3) 3 to 5 epochs. Table 6 compares the accuracy of various approaches; we report test set performance using the settings that corresponded to the best dev set performance.

4.3.1 Jointly Judge then Select (JJS)

In this scenario we sought to evaluate the QA component performance of six different models on the GSE1 subset (i.e., with gold SEs). The GSE1 column in Table 6 gives the test set accuracy rates of various approaches. As the *SJS* model has special handling for negation and “以上皆是 (all-of-the-above)” or “以上皆非 (none-of-the-above)” questions, it yields better performance than *BERT-MC Only* (0.862 vs. 0.849). The *SJCS* model further replaces the “以上皆是 (all-of-the-above)” and “以上皆非 (none-of-the-above)” options with the concatenation of the three other options. However, this degrades the baseline performance significantly, from 0.849 to 0.822. This is because the “以上皆是 (all-of-the-above)” and

Table 6. Test-set performance comparison.

	GSE1 [a]	***GSE1-Neg*** [b]	***GSE1-AllAbv&NonAbv*** [c]	***LSE*** [d]
BERT-MC only (baseline)	0.849	0.200	0.643	0.692
SJS	0.862	NA	NA	0.694
SJCS	0.822	NA	NA	0.661
BERT-MC ***+ Neg***	0.870	**0.400**	NA	0.695
BERT-MC ***+ AllAbv&NonAbv***	0.879	NA	**0.839**	0.719
BERT-MC ***+ Neg*** ***+ AllAbv&NonAbv*** ***(also JJS)***	**0.879**	NA	NA	**0.725**

[a] GSE1: SE1 subset with gold SEs.

[b] GSE1-Neg: Only negation-type questions within GSE1.

[c] GSE1-AllAbv&NonAbv: Only AllAbv&NonAbv-type questions within GSE1.

[d] LSE: <SE1+SE2> with all SEs retrieved from the Lucene search engine.

"以上皆非 (none-of-the-above)" options are closely related to the other three options. However, as it considers the concatenation option and the other three options independently, or separately, without using a complicated decision tree (specified in Figure 3), this approach is unable to take such correlation into account.

The *JJS* model (i.e., the last row in Table 6) addresses this problem by considering all of the options together simultaneously. Table 6 shows that it considerably outperforms the *SJCS* model by 5.7% (87.9% - 82.2%) on the test set, which shows that jointly processing all options together is essential after the concatenation step. The *BERT-MC+Neg* and *BERT-MC+AllAbv&NonAbv* models are also evaluated as an ablation analysis. Table 6 indicates they also outperform the *BERT-MC only* baseline by 2.1% (87.0% - 84.9%) and 3.0% (87.9% - 84.9%) on the test set, respectively, which shows the necessity of both the preprocessor and answer-selector modules.

Last, to explore the effects of the proposed approaches on specific question types, we conducted two additional experiments on two GSE1 subsets: (1) the *Neg-type only* subset, which contains only negation questions, to compare the performance between the BERT-MC only and *BERT-MC+Neg* approaches to evaluate the effectiveness of the answer-selector module; (2) the *AllAbv&NonAbv* only subset, which contains only *AllAbv* or *NonAbv* questions, to compare the *BERT-MC only* and *BERT-MC+AllAbv&NonAbv* approaches to evaluate the effectiveness of the proposed preprocessor. Table 6 clearly shows

Table 7. Error case of "BERT-MC+Neg" on "GSE1-Neg" subset.

SEs: 另外，隨著商業興盛，在府城、鹿港、艋舺等大城市，也出現由商人組成的「郊」。「郊」類似現代同業公會，成員除了經營貿易外，也積極參與地方的公共事務。

Question: 清朝統治臺灣時期，怎樣的人應該比較沒有共同的血緣？

Options: (1)參加同一個宗親會 (2)參加同一個祭祀公業 (3)參加同一個「郊」
(4) 在同一座宗祠祭祀祖先

Table 8. Error case of "BERT-MC+AllAbv&NonAbv" on "GSE1-AllAbv&NonAbv" subset.

SE:工業生產如果沒有適當處理，很容易破壞周遭環境，造成空氣汙染、噪音汙染、水質汙染、土地汙染等。例如：工業廢水或是家庭汙水直接排入河流，不僅危害河流生態，有毒物質如果流入大海，通過食物鏈進入人體，更會嚴重損害健康。

Question: "志忠家附近有一間工廠，時常將未經處理的汙水排入河川中，這樣可能會造成什麼後果？"

Options: (1)空氣汙染 (2)噪音 (3)水質汙染 (4)以上皆是

that the preprocessor (GSE1-Neg column) and answer-selector (GSE1-AllAbv&NonAbv column) modules effectively enhance *BERT-MC* on these two subsets (from 20% to 40%, and from 64.3% to 83.9%, respectively). The above experiments sufficiently demonstrate the effectiveness of our proposed approaches (unnecessary combinations are marked "NA" in Table 6).

The remaining errors in the *GSE1-Neg* and *GSE1-AllAbv&NonAbv* subsets are mainly due to that answering those questions requires further inference capability. Table 7 shows that we need to know that "商人 (businessmen)" are people without "共同的血緣 (blood relations)". Similarly, Table 8 shows that we need to know that "未經處理的汙水排入河川 (untreated sewage discharged into the river)" causes "水質汙染 (water pollution)".

4.3.2 LSE (SE1+SE2 with all SEs retrieved from Lucene)

Since the gold SE is not available for real-world applications, this scenario compares the system performance of different models in a real-world situation. That is, we evaluated various models with all the SEs retrieved from a search engine (i.e., *Apache Lucene* (https://lucene.apache.org/)). Furthermore, to support those questions for which no associated SEs from the lessons (i.e., the SE2 subset), we used Wikipedia as an external knowledge resource to provide SEs when possible. We first used Lucene to search the Taiwan elementary-school social studies textbook and Wikipedia separately to yield two different SEs, after which we constructed a fused SE by concatenating these two SEs with the format "**Textbook-SE [SEP] Wiki-SE**" where Textbook-SE and Wiki-SE denote the two SEs retrieved from the textbook and Wikipedia, respectively.

Table 9. Error types.

Error Type	Questions
Incorrect supporting evidence (52%)	**Wrong SE**: *清朝統治臺灣初期，漢人渡海來臺後，往往同鄉人聚居在一起，並且建築廟宇供奉共同信仰的神明。* **Question**: 臺灣有許多從中國移民來的漢人，來臺要渡過危險的臺灣海峽，所以什麼神明就被所有移民所共同信仰？ **Options**: (1)關公 (2)土地公 (3)媽祖 (4)三山國王
Requires advanced inference capability (48%)	**SE**: 刑法對傷害他人的行為加以處罰；民法則以損害賠償的方式，請問牛奶的保存期限過了沒？（相關法律：民法、消費者保護法、食品安全衛生管理法） **Question**: “小花在超市買到過期的餅乾，請問該超市的販售行為違反什麼法律？” **Options**: (1)刑法 (2)憲法 (3)教育基本法 (4)食品安全衛生管理法

Experimental results (the LSE column in Table 6) show that both the preprocessor and the answer selector effectively supplement BERT-MC; performance is improved further when they are jointly adopted (3.3% = 72.5% - 69.2%). Furthermore, the accuracy of the BERT-MC only model on LSE is significantly lower than that on GSE1 (69.2% vs. 84.9%), which clearly illustrates that extracting good SEs is essential in QA tasks. Last, to show the influence of incorporating Wikipedia, we conducted an experiment in which we used only Lucene to search the textbook. The BERT-MC+Neg+AllAbv&NonAbv model now drops to 70.4% (not shown in Table 6) from 72.5%, which shows that Wikipedia provides the required common sense for some cases.

5. Error Analysis and Discussion

We randomly selected 40 error cases from the test set of the *BERT-MC+Neg+AllAbv&NonAbv* model under the “*all SEs retrieved from Lucene*” scenario. We found that all errors come from two sources: (1) the correct support evidence was not retrieved (52%), and (2) the answer requires deep inference (48%). Table 9 shows an example for each category. For the first example, the retrieved SE is irrelevant to the question; our model thus fails to produce the correct answer. The second example illustrates that the model requires further inference capability to know that both “牛奶的保存期限過了沒 (Has the milk expired?)” and “在超市買到過期的餅乾 (I bought expired cookies in the supermarket)” are similar events related to “食品安全衛生管理法 (Act Governing Food Safety and Sanitation)”.

6. Related Work

Before 2015, most work on entailment judgment adopted statistical approaches (Kouylekov & Magnini, 2005; Heilman & Smith, 2010). In subsequent work, neural network models were widely adopted due to the availability of large datasets such as RACE (Lai *et al*., 2017) and SNLI (Bowman *et al*., 2015). Parikh *et al*. (2017) propose the first alignment-and-attention mechanism, achieving state-of-the-art (SOTA) results on the SNLI dataset. Chen *et al*. (2017) further propose a sequential inference model based on chain LSTMs which outperforms previous models. In recent work, pre-trained language models such as BERT (Devlin *et al*., 2019), XLNET (Yang *et al*., 2019), RoBERTa (Liu *et al*., 2019) and ALBERT (Lan *et al*., 2019) yield superior performance on MC RC tasks. However, these results are obtained mainly by utilizing surface features (Jiang & Marneffe, 2019). Besides, Zhang *et al*. (2020) propose a dual co-matching network to model relationships among passages, questions, and answer candidates to achieve SOTA results for MC questions. Also, Jin *et al*. (2020) propose two-stage transfer learning for coarse-tuning on out-of-domain datasets and fine-tuning on larger in-domain datasets to further improve performance. In comparison with those previous approaches, instead of adopting a new inference NN, our proposed approaches supplement the original BERT with additional modules to address two specific problems that BERT handles poorly.

7. Conclusion

We present several novel approaches to supplement BERT with additional modules to address problems with three specific types of questions that BERT-MC handles poorly (i.e., negation, all-of-the-above, and none-of-the-above). The proposed approach constitutes a new way to enhance a complicated DNN model with additional modules to pinpoint problems found in error analysis. Experimental results show the proposed approaches effectively improve performance, and thus demonstrate the feasibility of supplementing BERT with additional modules to fix specific problems.

Reference

Bowman, S. R., Angeli, G., Potts, C. & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632-642. https://doi.org/10.18653/v1/D15-1075

Chen, D. (2018). Neural Reading Comprehension and Beyond. (Doctoral Dissertation). Stanford Univ..

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H. & Inkpen, D. (2017). Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the*

Association for Computational Linguistics, 1657-1668. https://doi.org/10.18653/v1/P17-1152

Dagan, I., Glickman, O., & Magnini, B. (2005) The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, Springer,* 177-190. https://doi.org/10.1007/11736790_9

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171-4186. https://doi.org/10.18653/v1/N19-1423

Heilman, M. & Smith, N. A. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 1011-1019.

Jiang, N. & Marneffe, M.-C. D. (2019). Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 6086-6091. https://doi.org/10.18653/v1/D19-1630

Jin, D., Gao, S., Kao, J. Y., Chung, T., & Hakkani-tur, D. (2020). MMM: Multi-stage multi-task learning for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 8010–8017. https://doi.org/10.1609/aaai.v34i05.6310

Kouylekov, M. & Magnini, B. (2005). Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the First Challenge Workshop Recognising Textual Entailment 2005,* 17-20.

Lai, G., Xie, Q., Liu, H., Yang, Y. & Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 785–794. https://doi.org/10.18653/v1/D17-1082

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.

Lee, D., Liang, C. C. & Su, K. Y. (2020). Answering Chinese Elementary School Social Study Multiple Choice Questions. In *Proceedings of the 2020 International Conference on Technologies and Applications of Artificial Intelligence*.

Lin, Y.C. & Su, K.Y. (2021). How Fast can BERT Learn Simple Natural Language Inference? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 626-633. https://doi.org/10.18653/v1/2021.eacl-main.51

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Liu, S., Zhang, X., Zhang, S., Wang, H., & Zhang, W. (2019). Neural Machine Reading Comprehension: Methods and Trends. *Applied Sciences, 9*(18)*,* 3698. https://doi.org/10.3390/app9183698

Loshchilov, I. & Hutter, F. (2019). Decoupled Weight Decay Regularization. In Proceedings of *International Conference on Learning Representations 2019*.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). Doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano

Parikh, A. P., Tackstrom, O., Das, D. & Uszkoreit, J. (2016). A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 2249-2255. https://doi.org/10.18653/v1/D16-1244

Qiu, B., Chen, X., Xu, J., & Sun, Y. (2019). A Survey on Neural Machine Reading Comprehension. arXiv preprint arXiv:1906.03824.

Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 2383–2392. https://doi.org/10.18653/v1/D16-1264

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP,* 353–355. https://doi.org/10.18653/v1/W18-5446

Wu, T. M. & Su, K. Y. (2020). Making Negation-word Entailment Judgment via Supplementing BERT with Aggregative Pattern. In *International Conference on Technologies and Applications of Artificial Intelligence (TAAI 2020),* 17-22. https://doi.org/10.1109/TAAI51410.2020.00012

Xu, K., Tin, J., & Kim, J. (2019). A BERT based model for Multiple-Choice Reading Comprehension. Retrieved from http://cs229.stanford.edu/proj2019spr/report/72.pdf

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. & Le, Q. C. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of Advances in neural information processing systems 32 (NIPS 2019),* 5753–5763.

Zhang, S., Zhao, H., Wu, Y., Zhang, Z., Zhou, X., & Zhou, X. (2020). DCMN+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 9563-9570. https://doi.org/10.1609/aaai.v34i05.6502