Self-Teaching Machines to Read and Comprehend with Large-Scale Multi-Subject Question-Answering Data

Dian Yu[1] **Kai Sun**[2] **Dong Yu**[1] **Claire Cardie**[2]
[1]Tencent AI Lab, Bellevue, WA
[2]Cornell University, Ithaca, NY
{yudian, dyu}@tencent.com, ks985@cornell.edu, cardie@cs.cornell.edu

Abstract

Despite considerable progress, most machine reading comprehension (MRC) tasks still lack sufficient training data to fully exploit powerful deep neural network models with millions of parameters, and it is laborious, expensive, and time-consuming to create large-scale, high-quality MRC data through crowdsourcing. This paper focuses on generating more training data for MRC tasks by leveraging existing question-answering (QA) data. We first collect a large-scale multi-subject multiple-choice QA dataset for Chinese, ExamQA. We next use incomplete, yet relevant snippets returned by a web search engine as the context for each QA instance to convert it into a weakly-labeled MRC instance. To better use the weakly-labeled data to improve a target MRC task, we evaluate and compare several methods and further propose a self-teaching paradigm. Experimental results show that, upon state-of-the-art MRC baselines, we can obtain $+5.1\%$ in accuracy on a multiple-choice Chinese MRC dataset, C^3, and $+3.8\%$ in exact match on an extractive Chinese MRC dataset, CMRC 2018, demonstrating the usefulness of the generated QA-based weakly-labeled data for different types of MRC tasks as well as the effectiveness of self-teaching. ExamQA will be available at `https://dataset.org/examqa/`.

1 Introduction

Constructing high-quality, large-scale data remains a major challenge for machine reading comprehension (MRC) tasks, which aim to answer questions derived from a given document (Richardson et al., 2013; Hermann et al., 2015; Rodrigo et al., 2015). And it is laborious, expensive, and time-consuming to create large-scale MRC data through crowdsourcing, considering factors such as ensuring a high degree of difficulty for the questions and strong relevance between the designed questions and their associated documents. Therefore, crowdsourced MRC datasets, especially those requiring external knowledge beyond the given text (e.g., (Richardson et al., 2013; Ostermann et al., 2018; Huang et al., 2019a)), are usually small-scale, making it difficult to fully exploit prevailing MRC approaches based on pre-trained language models with millions of parameters (Devlin et al., 2019).

To alleviate this problem, most previous studies utilize the data of a target MRC task (Yang et al., 2017; Yu et al., 2018; Asai and Hajishirzi, 2020) or other MRC datasets of the same task type (Alberti et al., 2019) for data augmentation. In contrast, we examine the potential of using subject-area question answering data to generate additional MRC training data, motivated by the following two considerations. First, at some level, MRC and question answering (QA), which standardly requires retrieval of snippets of text from a large corpus that answer a given question (Voorhees and Tice, 2000; Burger et al., 2001; Fukumoto and Kato, 2001), seem to be quite related, and it has been demonstrated that medium-scale MRC datasets can be employed to improve performance of QA systems on small-scale subject-area QA datasets (Sun et al., 2019b; Pan et al., 2019). Second, there exists an enormous amount of real-world QA data across various subjects created by subject-matter experts, which is relatively easy and cheap to acquire but seldom used to help other tasks such as MRC.

As most of the existing multi-subject QA datasets are relatively small-scale, we first collect a large-scale **Q**uestion-**A**nswering dataset from **Exam**s (**ExamQA**) covering a wide range of subjects (e.g., sociology, education, and psychology), which contains 638k multiple-choice instances. We then present a method to convert QA instances in ExamQA into training instances for a target MRC task to benefit from knowledge transfer (Ruder et al., 2019). Unlike previous studies that augment each QA instance with relevant sentences retrieved from offline corpora, we rely on a standard

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 56–68
November 7–11, 2021. ©2021 Association for Computational Linguistics

information-seeking protocol enabled by modern search engines: users type their questions into a web search engine and read through the snippets from a variety of sources returned by the search engine to seek potential answers. Imitating this protocol, we use relevant snippets retrieved by a web search engine as the context of each QA instance. We regard such an MRC instance as weakly-labeled as the context is a form of distant supervision: while it might contain the answer to the question as required for MRC, it is also likely to be noisy, incomplete, and/or irrelevant (Section 3). Nevertheless, we find that this method for adding context to QA instances outperforms an approach that uses information from a single source such as Wikipedia as context for QA instances (Section 5.7).

There is also a challenge of using the large-scale QA-based weakly-labeled MRC data to improve a small-scale MRC task. We implement and compare several methods that use weakly-labeled data, such as classical sequential transfer learning (Ruder et al., 2019) and a very recent teacher-student paradigm with **multiple** teachers trained with different subsets of weakly-labeled data (Sun et al., 2020a) to generate soft-labeled MRC data for students. Furthermore, inspired by self-training (Yarowsky, 1995; Riloff, 1996) that iteratively regards the student as a teacher to relabel the **unlabeled** data for training a new student, we propose a paradigm, called **self-teaching**, to iteratively train a **single** teacher to provide soft labels of weakly-labeled or target MRC data. We always use the ground-truth hard labels of ExamQA to obtain more reliable soft labels of weakly-labeled data (Section 4). The "naturally" injected noise caused by context retrieval of our approach seems to help models learn better from weakly-labeled MRC data, playing a similar role as the noise (e.g., dropout and stochastic depth) that is intentionally injected into student models in previous studies (e.g., (He et al., 2020; Xie et al., 2020b)) (Section 5.7).

We study the effect of our large-scale weakly-labeled MRC data on representative MRC datasets for Chinese: a multiple-choice dataset, C^3 (Sun et al., 2020b), in which most questions cannot be solved solely by matching or paraphrasing, and an extractive dataset, CMRC 2018 (Cui et al., 2019), in which all answers are spans in the given documents. Experimental results show that soft-label paradigms such as multi-teacher and self-teaching achieve better performance than hard-label baselines. In particular, self-teaching does not need to carefully divide data for training several teachers at one stage as multi-teacher, yet performs equally well or better than multi-teacher. Based on state-of-the-art baselines (Xu et al., 2020; Cui et al., 2020), self-teaching leads to an $+5.1\%$ in accuracy on C^3 and $+3.8\%$ in exact match on CMRC 2018 over the same baselines without using any extra training data. We also demonstrate that our QA-based MRC data can be easily combined with other types of weakly-labeled MRC data in which noise is introduced by different factors (e.g., machine translation and knowledge extraction) for further gains (e.g., up to $+2.5\%$ in accuracy on C^3). As the proposed paradigm is language-independent and knowledge in many subjects (e.g., Mathematics and Physics) can also be culture-independent, we hope this work will benefit other tasks in different languages, perhaps through powerful multi-lingual language models or machine translation.

The contributions of this paper are as follows.

- We collect the largest multi-subject QA dataset to date to facilitate MRC/QA studies.
- Our study is the first to investigate the potential of using large-scale multi-subject QA data for MRC data augmentation.
- We evaluate and compare several methods to use the generated QA-based weakly-labeled MRC data. We further propose a simple yet effective self-teaching paradigm to better utilize large-scale weakly-labeled data.
- We show that our QA-based weakly-labeled MRC data can be easily used along with other types of weakly-labeled data for further gains.

2 Related Work

2.1 From Question Answering to Machine Reading Comprehension

This work is related to data augmentation in semi-supervised MRC studies, which partially or fully rely on the document-question-answer triples (Yang et al., 2017; Yuan et al., 2017; Yu et al., 2018; Zhang and Bansal, 2019; Zhu et al., 2019; Dong et al., 2019; Sun et al., 2019b; Alberti et al., 2019; Asai and Hajishirzi, 2020; Rennie et al., 2020) of target MRC tasks or at least similar domain corpora (Dhingra et al., 2018). We focus on leveraging multi-domain QA data to improve different types of general-domain MRC tasks.

2.2 Teacher-Student Paradigms

Teacher-student paradigms are widely used for knowledge distillation (Ba and Caruana, 2014; Li et al., 2014; Hinton et al., 2015). We aim to let a student model outperform its teacher model for performance improvements and thus use the same architecture for all teacher and student models.

Our work is related to self-training (Yarowsky, 1995; Riloff, 1996). The main differences are (i) noise is introduced by retrieved context instead of noisy answers, (ii) we generate weakly-labeled data based on existing large-scale QA data covering a wide range of domains, instead of the same domain (He et al., 2020; Xie et al., 2020a; Zhao et al., 2020; Chen et al., 2020) or at least approximately in-domain (Du et al., 2020) as the target MRC task, and (iii) ground-truth labels of weakly-labeled data are used directly or indirectly to train teacher models. Note that we use teacher models to generate new soft labels for fixed weakly-labeled data instead of new pseudo data with noisy labels from unlabeled data (e.g., (Wang et al., 2020a)).

Compared with previous multi-teacher student paradigms (You et al., 2019; Wang et al., 2020b; Yang et al., 2020), to train models to be strong teachers, we conduct iterative training and leverage large-scale weakly-labeled data rather than using clean, human-labeled data of similar tasks.

3 Weakly-Labeled Data Generation

3.1 Question-Answering Data Collection

We collect large-scale QA instances from freely accessible exams (including mock exams) designed for a variety of subjects such as programming, journalism, and ecology. We only keep multiple-choice single-answer instances written in Chinese. After deduplication, we obtain 638,436 QA instances.

To assess the subject coverage of ExamQA, we follow the subject list from China national standard (GB/T 13745-2009) (Standardization Administration of China, 2009) and check for each subject in the list if the name of the subject appears in the title of any exam to estimate the lower bound of subject coverage. The estimation shows that ExamQA covers **at least** 48 out of 62 first-level subjects and 187 out of 676 second-level subjects. Note that the actual subject coverage of ExamQA may be greatly underestimated, as only 24.2% of titles contain a subject name. Based on questions in ExamQA that could be linked to a subject, the top ten most frequent first-level subjects are Clinical Medicine (17.3%), Management (11.4%), Pharmacy (10.0%), Chinese Medicine and Chinese Materia Medica (8.0%), Psychology (7.3%), Law (5.2%), Economics (4.8%), Education (4.4%), Biology (3.6%), and Sociology (3.2%). See complete subject-wise frequencies in Appendix A.5.

We do not annotate a small subset of questions for human performance, as most of the subject-area questions are from higher education exams that require advanced domain knowledge.

3.2 Comparisons with Existing Subject-Area Question-Answering Datasets

Subject-area QA is an increasingly popular direction focusing on closing the performance gap between humans and machines in answering questions collected from real-world exams that are carefully designed by subject-matter experts. These tasks are mostly in multiple-choice forms. In Table 1, we list several representative subject-area multiple-choice QA datasets: NTCIR-11 QA-Lab (Shibuki et al., 2014), QS (Cheng et al., 2016), MCQA (Guo et al., 2017), ARC (Clark et al., 2018), GeoSQA (Huang et al., 2019b), HEAD-QA (Vilares and Gómez-Rodríguez, 2019), EXAMS (Hardalov et al., 2020), JEC-QA (Zhong et al., 2020), and MEDQA (Jin et al., 2020).

dataset	**# of subjects°**	**subjects**	**language**	**size**
QS	1	history	zh	0.6K
GeoSQA	1	geography	zh	4.1K
JEC-QA	1	legal	zh	26.4K
ARC	1	science	en	7.8K
QA-Lab	1	history	en/ja	0.3K
HEAD-QA	1	healthcare	en/es	6.8K
MEDQA	1	medical	en/zh	61.1K
MCQA	6	multi-subject	en/zh	14.4K
EXAMS	24	multi-subject	ar/bg/...	24.1K
ExamQA	**48**	multi-subject	zh	**638.4K**

Table 1: Representative subject-area QA datasets collected from exams (°: we report the number of subjects stated by previous studies and the number of first-level subjects in ExamQA; language code: ISO 639-1).

Some multiple-choice MRC datasets for Chinese such as C^3 are collected from language exams designed to test the reading comprehension ability of a human reader. To prevent data leakage, we **exclude** multiple-choice instances that have associated materials (e.g., a reference document), which have a setting like that of standard MRC.

3.3 Bringing Context to Question Answering

In this section, we present a method to convert QA instances into multiple-choice or extractive

MRC instances to make the resulting data and target MRC task in a similar format, which may benefit from knowledge transfer (Ruder et al., 2019).

Previous studies attempt to convert a multiple-choice subject-area QA task to a multiple-choice MRC task by retrieving relevant sentences for each question from a clean corpus to form a document. In contrast to relying on a fixed corpus, we retrieve the top-ranked snippets using a publicly available search engine. Specifically, we send each question to the search engine as the query and collect snippets from the first result page. Typically, we can collect ten snippets for each QA instance. Since all instances are freely accessible online, it is likely that a retrieved snippet merely contains the original QA instance rather than relevant context sufficient for answering the question. Therefore, we discard a snippet if more than one answer option appears as a substring in the snippet. We concatenate the remaining snippets into a document as the context of each QA instance. See data statistics of ExamQA and retrieved context in Table 2. Due to this construction method, it is likely that a document is noisy, incomplete, informal, and/or irrelevant. We provide sample instances in Table 3.

To convert these multiple-choice MRC instances into extractive ones, we remove the wrong answer options of each multiple-choice MRC instance and append the start offsets of the exact mention of the correct answer option in its associated document (we consider the first mention when multiple mentions exist). We remove instances in which correct answers are not mentioned in the documents.

metric	value
average # of answer options	4.0
average question length (in characters)	39.5
average answer option length (in characters)	6.7
average context length (in characters)	907.6
non-extractive correct answer option (%)	68.4
character vocabulary size	13,258

Table 2: Data statistics of ExamQA with context.

C1: 1. + b / b is equivalent to ((int) a) + (b / b), which can be obtained according to the priority of the processor. (Int) This is a forced type conversion. After the forced conversion ((int) a) is generally the double conversion to the int type, most platforms round to zero... 2./b, both sides of the division sign are doubletype , The result is also doubleType. That is 1.000000; integer. The first 5 is the int type, int... 3 .; a = 5.5; b = 2.5; c = (int) a + b / b; printf (:. Best answer: (int) a + b / b = 6, should be (int) a means round a, and round a is 5 (rounding cannot be used here, rounding is discarded, then b / b is 2.5 / 2.5, etc... 2019 July 25th, 2016-Analysis: The type of the value of the mixed expression is determined by the type with the highest precision in the expression, so it can be seen that option B can be excluded. Note that the result of b / b should be 1.00000, and (int) a is 5, and the result of the addition is still double...

Q1: Suppose a and b are double constants, a=5.5, and b=2.5, the value of the expression (int)a+b/b is ().

A. 5.500000.
B. 6.000000. ⋆
C. 6.500000.
D. 6.

C2: November 22, 2016 It can be seen that it is not a white box test case design method, so the correct answer to question (31) is B. Black box testing is also called functional testing, which is to detect whether each function can be used normally. At the test site, treat the program as... November 18, 2016 Black box testing technology is also called functional testing, which tests the external characteristics of the software without considering the internal structure and characteristics of the software. The main purpose of black box testing is to discover the following types of errors: Are there any errors... [Answer Analysis]...

Q2: Black box testing is also called functional testing, and black box testing cannot find ().

A. terminal error.
B. communication error.
C. interface error.
D. code redundancy. ⋆

C3: July 21, 2014-Friedman believes that the transmission variable of monetary policy should be (). Please help to give the correct answer and analysis, thank you! Reward: 0 answer bean Questioner: 00***42 Release time: 2014-07-21 View...

Q3: Friedman believes that the transmission variable of monetary policy should be ().

A. excess reserve.
B. interest rate.
C. currency supply. ⋆
D. base currency.

Table 3: English translation of sample instances in ExamQA with retrieved context (⋆: correct option).

4 Self-Teaching Paradigm

We will introduce a self-teaching paradigm to better leverage large-scale weakly-labeled MRC data to improve the performance of existing supervised methods on an MRC task of interest, which is relatively small-scale. Due to limited space, here we only discuss multiple-choice tasks and we leave the reformulation (e.g., soft labels and loss functions) for extractive MRC tasks in Appendix A.

4.1 Training a Junior Teacher

In previous teacher-student frameworks for domain/knowledge distillation (You et al., 2019; Wang et al., 2020b; Sun et al., 2020a), multiple teachers are trained using different data. However, it is difficult to divide the QA-based weakly-labeled data into subsets by subjects or fine-grained types of knowledge needed for answering questions. Instead, we simply train a junior teacher model using the combined human-annotated target MRC data and the weakly-labeled data, both with hard labels.

Let V denote a set of human-annotated training instances and W denote a set of weakly-labeled instances. For each instance $t \in V \cup W$, we let m_t denote its total number of answer options, and $\boldsymbol{h}^{(t)}$ be a one-hot (hard-label) vector such that $h_j^{(t)} = 1$ if the j-th answer option is labeled as correct. We train a single junior teacher model, denoted by $\mathcal{T}$, and optimize $\mathcal{T}$ by minimizing

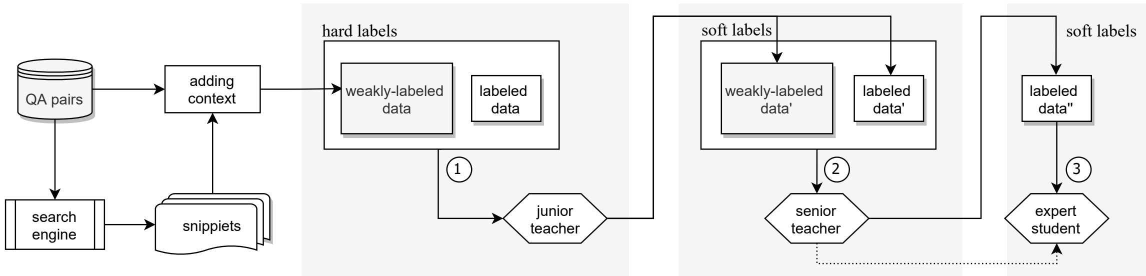

Figure 1: Self-teaching framework using large-scale QA data to improve relatively small-scale MRC.

$\sum_{t \in V \cup W} L_1(t, \theta_{\mathcal{T}})$; L_1 is defined as

$$L_1(t, \theta) = -\sum_{1 \leq k \leq m_t} h_k^{(t)} \log p_\theta(k \mid t),$$

where $p_\theta(k \mid t)$ denotes the probability that the k-th answer option of instance t is correct, estimated by the model with parameters θ.

4.2 Training a Senior Teacher

We then train a senior teacher model $\mathcal{S}$ using the same data as the junior teacher model $\mathcal{T}$ while replacing the hard labels of answer options with the soft labels predicted by $\mathcal{T}$ and the original hard labels. We define soft-label vector $\boldsymbol{s}^{(t)}$ for $t \in V \cup W$ such that

$$s_k^{(t)} = \lambda \, h_k^{(t)} + (1 - \lambda) p_{\theta_{\mathcal{T}}}(k \mid t),$$

where $\lambda \in [0, 1]$ is a weighting parameter, and $k = 1, \ldots, m_t$.

We optimize senior teacher $\mathcal{S}$ by minimizing $\sum_{t \in V \cup W} L_2(t, \theta_{\mathcal{S}})$, where L_2 is defined as

$$L_2(t, \theta) = -\sum_{1 \leq k \leq m_t} s_k^{(t)} \log p_\theta(k \mid t).$$

4.3 Training an Expert Student

As a final step, we initialize an expert student $\mathcal{E}$ with the resulting senior teacher model $\mathcal{S}$, and we fine-tune $\mathcal{E}$ on the target data V to help it achieve expertise in the task of interest, following most of the recent MRC methods (Radford et al., 2018; Devlin et al., 2019). This step differs from previous work in that we use the soft labels generated by the senior teacher model (Section 4.2) based on our assumption that a student model tends to learn better from a stronger teacher model. We will discuss more details in the experiment section and show that during self-training a student model tends to outperform its teacher model that provides soft labels to make itself a stronger teacher (Section 5).

We define new soft-label vector $\tilde{\boldsymbol{s}}^{(t)}$ for $t \in V$ such that

$$\tilde{s}_k^{(t)} = \lambda \, h_k^{(t)} + (1 - \lambda) p_{\theta_{\mathcal{S}}}(k \mid t),$$

where $\lambda \in [0, 1]$ is a weighting parameter, and $k = 1, \ldots, m_t$.

At this stage, we optimize $\mathcal{E}$ by minimizing $\sum_{t \in V} L_3(t, \theta_{\mathcal{E}})$, where L_3 is defined as

$$L_3(t, \theta) = -\sum_{1 \leq k \leq m_t} \tilde{s}_k^{(t)} \log p_\theta(k \mid t).$$

Figure 1 shows an overview of the proposed self-teaching paradigm.

4.4 Integrating Different Types of Weakly-Labeled MRC Data

We study the integration of multiple types of weakly-labeled data during weakly-supervised training with soft labels to save time and effort in retraining models on W with hard labels.

Take another weakly-labeled multiple-choice MRC data extracted automatically from television show and film scripts (Sun et al., 2020a) as an example, denoted as W_s, besides the weakly-labeled data W constructed based on existing QA instances. Following the above three-step procedure, we first train a junior teacher $\mathcal{T}_s$ using W_s to generate soft labels of W_s and V. We then train a senior teacher $\mathcal{S}_*$ upon the combination of soft-labeled W_s, W (Section 4.2), and V. Note that we simply use two versions of soft-labeled V generated by $\mathcal{T}$ and $\mathcal{T}_s$, respectively. The resulting senior teacher $\mathcal{S}_*$ is used to generate the final soft labels of V for training an expert student. In Section 5.6, we will discuss integration with other types of weakly-labeled MRC data in which the source of noise varies.

5 Experiments

5.1 Data Statistics

See statistics of two relatively small-scale MRC datasets ($C^3$and CMRC 2018) and three kinds of large-scale weakly-labeled MRC data in Table 4.

For CMRC 2018, we use its publicly available training and development sets. For weakly-labeled MRC data, besides the automatically extracted SCRIPT (Section 4.4), we also consider human-labeled multiple-choice MRC instances in other resource-rich languages such as English. We use Google Translate to translate instances from C^3's English counterparts RACE (Lai et al., 2017) and DREAM (Sun et al., 2019a) that are also collected from language exams into Chinese (referred to as MRC_{MT}).

MRC data	source	noise	# instances
human-annotated:			
C^3	language exams	–	19,577
CMRC 2018	Wikipedia	–	19,071
weakly-labeled:			
MRC_{MT}	language exams	translation	107,884
SCRIPT	TV/movie scripts	extraction	700,816
ExamQA	multi-subject exams	retrieval	638,436

Table 4: Human-annotated and weakly-labeled machine reading comprehension data statistics.

5.2 Implementation Details

We use Baidu Search to form a document for each QA instance. We follow recent state-of-the-art MRC methods for the model architecture that consists of a pre-trained language model and a classification layer. We use the same architecture for baselines and all teacher or student models. We use RoBERTa-wwm-ext-large (Cui et al., 2020) as the pre-trained language model for Chinese, which reaches state-of-the-art performance on representative MRC tasks for Chinese such as C^3 and CMRC 2018 (Xu et al., 2020). We are aware of the emerging newly-released pre-trained language models for Chinese and leave the exploration of them for future studies. We train a junior/senior teacher model for one epoch as large-scale weakly-labeled data is used. We train baselines and expert students for eight epochs on C^3 and two epochs on CMRC 2018. More epochs do not lead to better results on both MRC datasets. In all experiments, we set λ (defined in Section 4.2-4.3) to 0.5 to permit easy comparisons with the multi-teacher paradigm (Sun et al., 2020a) (Section 5.5), and we report the average score of five runs with different random seeds and standard deviation in brackets. See more setting details in Appendix A.4.

5.3 Main Results

In Table 5, for fair comparisons, we mainly compare methods built on the **same** pre-trained language model on the multiple-choice MRC dataset C^3. Under the zero-shot scenario using ExamQA, we already see promising results (e.g., 64.9% on the C^3 dev set). With the proposed self-teaching paradigm, expert student (4) improves baseline (1) based on the same model architecture by up to 5.1% in accuracy, and it outperforms two-stage fine-tuning (G) and sequential transfer learning (D). The two hard-label methods (the only difference lies in whether or not the target MRC training data is used at the first stage) are moderately effective but more efficient as weakly-labeled data is only used once. We will thoroughly compare self-teaching and the multi-teacher paradigm (Sun et al., 2020a) that also uses soft labels and weakly-labeled MRC data in different settings in Section 5.5.

For an extractive MRC task, we follow the self-teaching paradigm (Section 4) and introduce how to apply self-teaching to extractive tasks by redefining hard and soft labels for probability distributions of being answer start and end tokens, changing the loss function for senior teacher and expert student, etc., in Appendix A. As there are major differences (e.g., type of questions/answers and required prior knowledge) between extractive and multiple-choice MRC tasks, we do not see positive results by adapting the resulting best-performing expert student ((4) in Table 5) to initialize an extractive model.

As shown in Table 6, similarly, the expert student also reaches the best performance, outperforming the baseline model (Cui et al., 2020) implemented based on the same pre-trained language model by 3.8% in exact match and 2.0% in F1. As each (question, document) corresponds to two probability distributions in a much larger dimension compared to that of soft labels for multiple-choice tasks, due to memory limitations, we only use one third of the weakly-labeled extractive MRC data.

5.4 Observations and Discussions

Hereafter, we concentrate on multiple-choice tasks as we can afford to use more weakly-labeled MRC data, especially soft-labeled, during training. We compare our methods and other baselines in Table 5, and we have the following observations.

I. Under the self-teaching paradigm, student models tend to outperform their corresponding teacher models. For example, on the C^3 dataset, the accuracy of the senior teacher (3) is 1.5% higher than the result 75.6% achieved by its teacher (2).

II. Using a strong teacher model to provide

id	model	init.	teacher	training data name	training data label (H/S)	dev	test
A	AMBERT (Zhang and Li, 2020)	–	–	◇	H	69.5	69.6
B	ERNIE 2.0 (Sun et al., 2020c; Ding et al., 2021)	–	–	◇	H	72.3	73.2
C	RoBERTa-wwm-ext-large (Cui et al., 2020)	–	–	◇	H	–	73.8
D	sequential transfer learning (Ruder et al., 2019)*	–	–	1st: ExamQA; 2nd: ◇	H; H	76.3 (0.4)	76.1 (0.3)
E	two-stage fine-tuning (Sun et al., 2020a)	–	–	1st: ◇ + SCRIPT; 2nd: ◇	H; H	75.6	75.2
F	multi-teacher (Sun et al., 2020a)	–	–	1st/2nd: ◇ + SCRIPT; 3rd: ◇	H; S; S	77.4	77.7
	self-teaching:						
1	baseline (our implementation of C)	–	–	◇	H	73.9 (0.5)	73.4 (0.5)
2	junior teacher	–	–	◇ + ExamQA	H	74.0 (0.8)	75.6 (0.5)
3	senior teacher	–	2	◇ + ExamQA	S	75.7 (0.5)	77.1 (0.4)
4	expert student	3	3	◇	S	**78.2** (0.3)	**78.5** (0.2)
	other expert variants or baselines:						
5	expert student (weak teacher)	3	2	◇	S	77.8 (0.4)	78.0 (0.3)
6	expert student (weak initialization)	–	3	◇	S	74.9 (0.3)	74.8 (0.5)
G	two-stage fine-tuning (same as E)	2	–	◇	H	76.5 (0.3)	76.6 (0.8)
H	basic teacher-student w/o ExamQA	1	1	◇	S	73.4 (0.4)	72.6 (0.4)

Table 5: Average accuracy and standard deviation (%) on the dev and test sets of the C^3 dataset (H/S: hard/soft; $\star$: our implementations). ◇ is the training set of C^3 for all experiments; init. means the starting point, and – in this column means using the pre-trained language model for initialization.

model	extra training data	EM	F1
AMBERT	N/A	68.8	87.3
ERNIE 2.0	N/A	71.5	89.9
(Cui et al., 2020)	N/A	67.6	87.9
transfer learning	⋄	72.1 (0.6)	90.1 (0.3)
two-stage fine-tuning	⋄	71.4 (0.2)	89.8 (1.0)
baseline	N/A	70.3 (1.4)	89.2 (0.2)
junior teacher	⋄	71.8 (0.6)	89.8 (0.4)
senior teacher	⋄	72.5 (0.6)	90.1 (0.5)
expert student	N/A	**74.1** (0.7)	**91.2** (0.3)

Table 6: EM and F1 (%) on the publicly available development set of CMRC 2018 (⋄: subset of ExamQA used for training junior/senior teacher models).

soft labels helps across settings. We consider a teacher model to be strong if it achieves good performance on the target MRC task. Using the senior teacher (3), which is stronger than the junior teacher (2), to provide soft labels of C^3 to train an expert student results in +0.5% in accuracy ((4) vs. (5)). To explore whether this also applies to expert models, we experiment with a variant of expert student (4): still starting from the same senior teacher (3), we now put back expert student (4) as the teacher model to generate soft labels of C^3 to train a new expert student. However, this variant does not yield further gains (78.2 (0.4)) on the development set). Seeing more data than the expert student may make the more *"knowledgable"* senior teacher a better teacher to provide soft labels of the target MRC data. While it is possible to use the senior teacher itself to obtain a stronger senior teacher just as traditional self-training, it is much less efficient to retrain a model upon the large-scale weakly-labeled data than the above variant, which could be explored in future work.

III. Large-scale QA-based weakly-labeled data can be helpful for MRC. Using a basic teacher-student paradigm over the target MRC task alone even hurts the performance ((1) vs. (H) in Table 5). Under the self-training paradigm, helping train teacher models, especially the senior teacher that is further used as a good starting point of the expert student ((4) vs. (6)), reflects the usefulness of the large-scale weakly-labeled data. To train an expert student, we observe that both soft labels provided by a strong teacher and using the teacher for model initialization are necessary, as training the expert student from a pre-trained language model does not fully leverage the strength of the weakly-labeled data (e.g., (3) vs. (6)).

IV. Initializing a student with its teacher is not always useful. Though starting from the junior teacher slightly boosts (+0.3% in accuracy) a senior teacher's performance, using the resulting senior teacher to initialize and teach the expert student actually hurts performance (−0.7% in accuracy on the dev set). It is perhaps due to convergence of the junior teacher and senior teacher, which are already trained upon the same set of large-scale training data, although the labels are hard and soft, respectively. Similar observations have also been made in previous vision studies. For example, Xie et al. (2020b) reported that it is sometimes better to train a student from scratch than initializing the student with its teacher when large-scale pseudo-labeled data is consistently involved. Therefore, we do not use the junior teacher to initialize the senior teacher in our main experiment (3 in Table 5 and senior teacher in Table 6).

paradigm	weakly-labeled data	data segmentation criteria	# of junior teachers	dev	test
self-teaching	ExamQA	–	1	**78.2** (0.3)	**78.5** (0.2)
multi-teacher (our implementation)	ExamQA	random	4	77.5 (0.5)	77.9 (0.2)
self-teaching	SCRIPT	–	1	77.9 (0.4)	77.9 (0.4)
multi-teacher (our implementation)	SCRIPT	random	4	77.7 (0.2)	77.5 (0.3)
multi-teacher (our implementation)	SCRIPT	knowledge type	4	77.7 (0.4)	77.9 (0.3)

Table 7: Comparison of self-teaching and multi-teacher using different types of weakly-labeled data in accuracy (%) on the dev and test sets of the C^3 dataset.

5.5 Comparing Self-Teaching and Multi-Teacher Paradigms

Recent work (Sun et al., 2020a) shows that it is better to train multiple teacher models upon different subsets of weakly-labeled data with hard labels and then use these teachers to generate **soft** labels for both the weakly-labeled data and the small-scale MRC data for two-stage soft-label fine-tuning, compared against two-stage hard-label fine-tuning (i.e., (E) vs. (F) in Table 5). However, herein lies an unanswered question: **whether teacher models' data diversity or number matters to the resulting expert student's performance**.

As it is difficult to divide ExamQA into subsets by subjects, which can result in hundreds of teachers, we shuffle ExamQA and divide it into four subsets of similar size and follow the multi-teacher paradigm mentioned above. We find that self-teaching provides larger accuracy gains compared against multi-teacher when knowledge/domain-based data segmentation is tricky (Table 7).

We also consider the setting when it is easy to split data into subsets by the type of knowledge: we compare self-teaching with multi-teacher given the weakly-labeled data based on SCRIPT, which contains four subsets of verbal-nonverbal knowledge extracted by different patterns. Results show that self-teaching has competitive performance compared with multi-teacher that carefully feed different types of knowledge into different teachers, indicating that the impact of the number of teacher models may be **limited**. To further study the impact of data diversity of teachers, we shuffle SCRIPT and divide it into four subsets of similar size to train four teacher models. Using the same multi-teacher paradigm, we experimentally demonstrate a **weak** correlation between the data diversity of teachers and the final performance of the expert student.

5.6 Using ExamQA along with Other Types of Weakly-Labeled Data

Using the method mentioned in Section 4.4, introducing additional weakly-labeled MRC instances generated based on verbal-nonverbal knowledge automatically extracted from scripts, we observe $+1.5\%$ in accuracy over the best-performing expert student (4 in Table 5), which already outperforms the expert student obtained when we only use one-third of weakly-labeled data constructed based on ExamQA by 0.8% in accuracy (Table 8). Furthermore, we show it is possible to use the same procedure to adapt self-teaching to incorporate **extra noisy human-labeled** multiple-choice MRC instances (MRC_{MT} in this work), and we apply self-teaching to additionally incorporate the data, leading to $+2.5\%$ in accuracy. We do not study how to further improve machine reading comprehension by just using extra **clean** human-annotated MRC data, which is not the main focus of this paper. These results suggest the flexibility and scalability of self-teaching, and our QA-based weakly-labeled MRC data can be used with other types of weakly-labeled MRC data to further boost performance.

weakly-labeled MRC data	size	dev	test
–	–	73.9 (0.5)	73.4 (0.5)
subset of ExamQA	0.2M	77.8 (0.2)	77.7 (0.1)
ExamQA	0.6M	78.2 (0.3)	78.5 (0.2)
ExamQA + SCRIPT	1.3M	79.5 (0.2)	80.0 (0.2)
mixed-labeled data			
ExamQA + MRC_{MT}	0.7M	**80.4** (0.1)	**81.0** (0.2)

Table 8: Accuracy comparison of expert students, which are obtained when different size of weakly-labeled data is used during self-teaching, on the dev and test sets of the C^3 dataset (size: number of instances).

5.7 The Roles of Noise and Source of Context in Weakly-Labeled Data

As context returned by a web search engine is likely to be noisy, we conduct a preliminary experiment to evaluate the impact of noise in context by removing wrong answer options from the context of

source of context	denoise	dev	test
search engine	×	**78.2** (0.3)	**78.5** (0.2)
search engine	✓	77.0 (0.3)	77.5 (0.3)
Wikipedia	×	77.1 (0.3)	77.4 (0.2)

Table 9: Accuracy comparison of expert students on the dev and test sets of the C^3 dataset, which are obtained when different types of sources are used to form context of weakly-labeled data.

each weakly-labeled MRC instance. Surprisingly, context cleaning **hurts** accuracy by 1.2% on the development set of C^3. It is possible that noisy context helps improve the generalization ability of both teacher and student models, just as the noise that is intentionally added in previous work (e.g., (He et al., 2020; Xie et al., 2020b)).

Besides using snippets retrieved from a search engine to form context, we use the default search engine in Wikipedia to collect relevant snippets from Wikipedia for each question, leading to decreased accuracy (-1.1% on C^3), perhaps due to questions in ExamQA requires fine-grained subject-specific knowledge that is not always covered in Wikipedia articles written in Chinese.

6 Conclusions

We focus on using multi-subject QA instances to construct large-scale weakly-labeled MRC data to improve a target MRC task, which lacks sufficient training data. We collect a large-scale multi-subject multiple-choice QA dataset ExamQA and use incomplete, yet relevant snippets returned by a search engine as context of each QA instance to convert it into a weakly-labeled MRC instance. We evaluate and compare several methods and further propose self-teaching to better use these weakly-labeled MRC instances. Experimental results show that we can obtain $+5.1\%$ in accuracy on a multiple-choice MRC dataset C^3 and $+3.8\%$ in exact match on an extractive MRC dataset CMRC 2018, supporting the effectiveness of self-teaching and the usefulness of QA-based augmented data for MRC.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful and insightful feedback. The first author would also like to thank Qi Luan for productive discussions on Section 4.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the ACL*, pages 6168–6173.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the ACL*, pages 5642–5650.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Proceedings of the NIPS*, pages 2654–2662.

John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, et al. 2001. Issues, tasks and program structures to roadmap research in question & answering (q&a). In *Document Understanding Conferences Roadmapping Documents*, pages 1–35.

Yanda Chen, Md Arafat Sultan, and Vittorio Castelli. 2020. Improved synthetic training for reading comprehension. *arXiv preprint*, cs.CL/2010.12776v1.

Gong Cheng, Weixi Zhu, Ziwei Wang, Jianghui Chen, and Yuzhong Qu. 2016. Taking up the gaokao challenge: An information retrieval approach. In *Proceedings of the IJCAI*, pages 2479–2485.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint*, cs.CL/1803.05457v1.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the EMNLP*, pages 657–668.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the EMNLP-IJCNLP*, pages 5886–5891.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*, pages 4171–4186.

Bhuwan Dhingra, Danish Pruthi, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the NAACL-HLT*, pages 582–587.

Siyu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-Doc: A retrospective long-document modeling transformer. In *Proceedings of the ACL-IJCNLP*, pages 2914–2927.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the NeurIPS*, pages 13063–13075.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *arXiv preprint*, cs.CL/2010.02194v1.

Jun-ichi Fukumoto and Tsuneaki Kato. 2001. An overview of question and answering challenge (QAC) of the next NTCIR workshop. In *NTCIR*.

Shangmin Guo, Kang Liu, Shizhu He, Cao Liu, Jun Zhao, and Zhuoyu Wei. 2017. IJCNLP-2017 task 5: Multi-choice question answering in examinations. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 34–40.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the EMNLP*, pages 5427–5444.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *Proceedings of the ICLR*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the NIPS*, pages 1693–1701.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint*, stat.ML/1503.02531v1.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019a. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the EMNLP-IJCNLP*, pages 2391–2401.

Zixian Huang, Yulin Shen, Xiao Li, Yuang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, and Yuzhong Qu. 2019b. GeoSQA: A benchmark for scenario-based question answering in the geography domain at high school level. In *Proceedings of the EMNLP-IJCNLP*, pages 5865–5870.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arXiv preprint*, cs.CL/2009.13081v1.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the EMNLP*, pages 785–794.

Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. 2014. Learning small-size DNN with output-distribution-based criteria. In *Proceedings of the Interspeech*, pages 1910–1914.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the SemEval*, pages 747–757.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving question answering with external knowledge. In *Proceedings of the MRQA*, pages 27–37.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

Steven Rennie, Etienne Marcheret, Neil Mallinar, David Nahamoo, and Vaibhava Goel. 2020. Unsupervised adaptation of question answering systems via generative self-training. In *Proceedings of the EMNLP*, pages 1148–1157.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the EMNLP*, pages 193–203.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the AAAI*, pages 1044–1049.

Alvaro Rodrigo, Anselmo Penas, Yusuke Miyao, Eduard H Hovy, and Noriko Kando. 2015. Overview of CLEF QA entrance exams task 2015. In *CLEF (Working Notes)*.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the NAACL-HLT: Tutorials*, pages 15–18.

Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab task. In *Ntcir*.

Standardization Administration of China. 2009. The people's republic of China national standard (GB/T 13745-2009): Classification and code of disciplines. *Chinese Standard Publishing House, Beijing*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, and Claire Cardie. 2020a. Improving machine reading comprehension with contextualized commonsense knowledge. *arXiv preprint*, cs.CL/2009.05831v2.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association of Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019b. Improving machine reading comprehension with general reading strategies. In *Proceedings of the NAACL-HLT*, pages 2633–2643.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020b. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020c. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI*, volume 34, pages 8968–8975.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the ACL*, pages 960–966.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Shaolei Wang, Zhongyuan Wang, Wanxiang Che, and Ting Liu. 2020a. Combining self-training and self-supervised learning for unsupervised disfluency detection. In *Proceedings of the EMNLP*, pages 1813–1822.

Yong Wang, Longyue Wang, Shuming Shi, Victor OK Li, and Zhaopeng Tu. 2020b. Go from the general to the particular: Multi-domain translation with domain transformation networks. In *Proceedings of the AAAI*, pages 9233–9241.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020a. Unsupervised data augmentation for consistency training. In *Proceedings of the NeurIPS*, pages 6256–6268.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020b. Self-training with noisy student improves imagenet classification. In *Proceedings of the CVPR*, pages 10687–10698.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the COLING*, pages 4762–4772.

Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the WSDM*, pages 690–698.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the ACL*, pages 1040–1050.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the ACL*, pages 189–196.

Zhao You, Dan Su, and Dong Yu. 2019. Teach an all-rounder with experts in different domains. In *Proceedings of the ICASSP*, pages 6425–6429.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the ICLR*.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the RepL4NLP*, pages 15–25.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the EMNLP-IJCNLP*, pages 2495–2509.

Xinsong Zhang and Hang Li. 2020. Ambert: A pre-trained language model with multi-grained tokenization. *arXiv preprint*, cs.CL/2008.11869v3.

Zhenyu Zhao, Shuangzhi Wu, Muyun Yang, Kehai Chen, and Tiejun Zhao. 2020. Robust machine reading comprehension by learning soft labels. In *Proceedings of the COLING*, pages 2754–2759.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *Proceedings of the AAAI*, pages 9701–9708.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the ACL*, pages 4238–4248.

A Appendix

A.1 Training a Junior Teacher

Let V denote a set of human-labeled instances and W denote a set of weakly-labeled instances. Each instance contains a document d, a question q, and an answer span a in d. Let a_{start} and a_{end} denote, respectively, the start offset and end offset of a, which appears in d. For each instance t = (d, q, a), let l_t denote the length of the concatenated (q, d) taken as the input to an MRC model. We train a junior teacher model, denoted by $\mathcal{T}$, which learns to predict the probability of each token in the input to be the start or end token of the correct answer. Let $p_{\text{start},\theta}(k \mid t)$ and $p_{\text{end},\theta}(k \mid t)$ denote the probabilities that the k-th token in (q, d) to be the start and end token respectively, estimated by a model with parameters θ. We optimize $\mathcal{T}$ by minimizing $\sum_{t \in V \cup W} L_1(t, \theta_{\mathcal{T}})$, where L_1 is defined as

$$L_1(t, \theta) = -\log p_{\text{start},\theta}(a_{\text{start}} \mid t) - \log p_{\text{end},\theta}(a_{\text{end}} \mid t).$$

A.2 Training a Senior Teacher

We then train a senior teacher model $\mathcal{S}$ using the same data as the junior teacher model $\mathcal{T}$ while replacing the hard labels with the soft labels predicted by $\mathcal{T}$. We define $\boldsymbol{h}_{\text{start}}^{(t)}$ and $\boldsymbol{h}_{\text{end}}^{(t)}$ to be one-hot hard-label vectors such that $\boldsymbol{h}_{\text{start},i}^{(t)} = 1$ and $\boldsymbol{h}_{\text{end},j}^{(t)} = 1$ if the i-th and j-th tokens in (q, d) are the start and end token of the correct answer respectively. We define soft-label vectors $\boldsymbol{s}_{\text{start}}{}^{(t)}$ and $\boldsymbol{s}_{\text{end}}{}^{(t)}$ for $t \in V \cup W$ such that

$$\boldsymbol{s}_{\text{start},k}^{(t)} = \lambda\, \boldsymbol{h}_{\text{start},k}^{(t)} + (1 - \lambda) p_{\text{start},\theta_{\mathcal{T}}}(k \mid t)$$

and

$$\boldsymbol{s}_{\text{end},k}^{(t)} = \lambda\, \boldsymbol{h}_{\text{end},k}^{(t)} + (1 - \lambda) p_{\text{end},\theta_{\mathcal{T}}}(k \mid t),$$

where $\lambda \in [0, 1]$ is a weighting parameter, and $k = 1, \ldots, l_t$. We optimize senior teacher $\mathcal{S}$ by minimizing $\sum_{t \in V \cup W} L_2(t, \theta_{\mathcal{S}})$, where L_2 is defined as

$$L_{\text{start},2}(t, \theta) = -\sum_{1 \le k \le l_t} \boldsymbol{s}_{\text{start},k}^{(t)} \log p_{\text{start},\theta}(k \mid t)$$

$$L_{\text{end},2}(t, \theta) = -\sum_{1 \le k \le l_t} \boldsymbol{s}_{\text{end},k}^{(t)} \log p_{\text{end},\theta}(k \mid t)$$

$$L_2(t, \theta) = \frac{1}{2}(L_{\text{start},2}(t, \theta) + L_{\text{end},2}(t, \theta)).$$

A.3 Training an Expert Student

We now introduce the formulation of training expert student $\mathcal{E}$. For instance $t \in V$, we define new soft-label vectors $\tilde{\boldsymbol{s}}_{\text{start}}^{(t)}$ and $\tilde{\boldsymbol{s}}_{\text{end}}^{(t)}$ such that

$$\tilde{\boldsymbol{s}}_{\text{start},k}^{(t)} = \lambda\, \boldsymbol{h}_{\text{start},k}^{(t)} + (1 - \lambda) p_{\text{start},\theta_{\mathcal{S}}}(k \mid t)$$

and

$$\tilde{\boldsymbol{s}}_{\text{end},k}^{(t)} = \lambda\, \boldsymbol{h}_{\text{end},k}^{(t)} + (1 - \lambda) p_{\text{end},\theta_{\mathcal{S}}}(k \mid t),$$

where $\lambda \in [0, 1]$ is a weighting parameter, and $k = 1, \ldots, l_t$. We optimize $\mathcal{E}$ by minimizing $\sum_{t \in V} L_3(t, \theta_{\mathcal{E}})$, where L_3 is defined as

$$L_{\text{start},3}(t, \theta) = -\sum_{1 \le k \le l_t} \tilde{\boldsymbol{s}}_{\text{start},k}^{(t)} \log p_{\text{start},\theta}(k \mid t)$$

$$L_{\text{end},3}(t, \theta) = -\sum_{1 \le k \le l_t} \tilde{\boldsymbol{s}}_{\text{end},k}^{(t)} \log p_{\text{end},\theta}(k \mid t)$$

$$L_3(t, \theta) = \frac{1}{2}(L_{\text{start}}(t, \theta) + L_{\text{end}}(t, \theta)).$$

A.4 Settings

	jt/st	**es/baseline**
training data	ExamQA + C^3	C^3
initial learning rate	2e-5	2e-5
batch size	24	24
# of training epochs	1	8
max sequence length	512	512
training labels	hard/soft	soft/hard

Table 10: Hyper-parameter settings for training multiple-choice machine reading comprehension models (jt: junior teacher; st: senior teacher; es: expert student).

	jt/st	**es/baseline**
training data	⋄ + CMRC 2018	CMRC 2018
initial learning rate	3e-5	3e-5
batch size	32	32
# of training epochs	1	2
max sequence length	512	512
training labels	hard/soft	soft/hard

Table 11: Hyper-parameter settings for training extractive machine reading comprehension models (⋄: subset of ExamQA; jt: junior teacher; st: senior teacher; es: expert student).

A.5 Subjects in ExamQA

subject id	subject name	name translation	# of questions
110	数学	Mathematics	2,875
120	信息科学与系统科学	Information Science and System Science	6
130	力学	Mechanics	1,354
140	物理学	Physics	606
150	化学	Chemistry	3,634
170	地球科学	Earth Science	131
180	生物学	Biology	6,554
190	心理学	Psychology	13,317
210	农学	Agronomy	523
230	畜牧、兽医科学	Animal Husbandry and Veterinary Science	98
310	基础医学	Basic Medicine	5,526
320	临床医学	Clinical Medicine	31,412
330	预防医学与公共卫生学	Preventive Medicine and Public Health	1,132
350	药学	Pharmacy	18,171
360	中医学与中药学	Chinese Medicine and Chinese Materia Medica	14,470
413	信息与系统科学相关工程与技术	Information and System Science Related Engineering and Technology	140
416	自然科学相关工程与技术	Natural Science Related Engineering and Technology	14
420	测绘科学技术	Surveying and Mapping Science and Technology	31
430	材料科学	Materials Science	107
460	机械工程	Mechanical Engineering	348
470	动力与电气工程	Power and Electrical Engineering	2,438
510	电子与通信技术	Electronics and Communications Technology	945
520	计算机科学技术	Computer Science and Technology	4,867
530	化学工程	Chemical Engineering	156
550	食品科学技术	Food Science and Technology	28
560	土木建筑工程	Civil Engineering	1,660
570	水利工程	Water Conservancy Engineering	270
580	交通运输工程	Transportation Engineering	833
610	环境科学技术及资源科学技术	Environmental/Resource Science and Technology	23
620	安全科学技术	Safety Science and Technology	49
630	管理学	Management	20,771
710	马克思主义	Marxism	1,225
720	哲学	Philosophy	1,629
730	宗教学	Religious Studies	34
740	语言学	Linguistics	113
750	文学	Literature	3,806
760	艺术学	Art	3,423
770	历史学	History	1,387
790	经济学	Economics	8,784
810	政治学	Political Science	3,996
820	法学	Law	9,442
840	社会学	Sociology	5,802
850	民族学与文化学	Ethnology and Cultural Studies	15
860	新闻学与传播学	Journalism and Communication	858
870	图书馆、情报与文献学	Library, Information, and Documentation	144
880	教育学	Education	8,002
890	体育科学	Sports Science	49
910	统计学	Statistics	546
–	–	Unclassified	456,692

Table 12: Subject-wise frequencies of questions in ExamQA.