Improving Machine Reading Comprehension with Contextualized Commonsense Knowledge

Kai Sun[1†]   **Dian Yu**[2†]   **Jianshu Chen**[2]   **Dong Yu**[2]   **Claire Cardie**[1]

[1]Cornell University, Ithaca, NY
[2]Tencent AI Lab, Bellevue, WA
ks985@cornell.edu, {yudian, jianshuchen, dyu}@tencent.com, cardie@cs.cornell.edu

Abstract

In this paper, we aim to extract commonsense knowledge to improve machine reading comprehension. We propose to represent relations implicitly by situating structured knowledge in a context instead of relying on a predefined set of relations, and we call it contextualized knowledge. Each piece of contextualized knowledge consists of a pair of interrelated verbal and nonverbal messages extracted from a script and the scene in which they occur as context to implicitly represent the relation between the verbal and nonverbal messages, which are originally conveyed by different modalities within the script. We propose a two-stage fine-tuning strategy to use the large-scale weakly-labeled data based on a single type of contextualized knowledge and employ a teacher-student paradigm to inject multiple types of contextualized knowledge into a student machine reader. Experimental results demonstrate that our method outperforms a state-of-the-art baseline by a 4.3% improvement in accuracy on the machine reading comprehension dataset C^3, wherein most of the questions require unstated prior knowledge.

1 Introduction

During the past few years, there is a trend of taking advantage of existing commonsense knowledge graphs such as ConceptNet (Speer et al., 2017) or automatically constructed graphs (Zhang et al., 2020) to improve machine reading comprehension (MRC) tasks that contain a high percentage of questions requiring commonsense knowledge unstated in the given documents (Mostafazadeh et al., 2016; Lai et al., 2017; Ostermann et al., 2018; Sun et al., 2019a; Huang et al., 2019). In this paper, following

the second line of work, we aim to extract commonsense knowledge from external unstructured corpora and explore using the structured knowledge to improve machine reading comprehension.

Typically, each piece of commonsense knowledge is represented as a triple that contains two phrases (e.g., (*"finding a lost item"*, *"happiness"*) and the relation (e.g., CAUSES) between phrases, which can be one of a small pre-defined set of relations (Tandon et al., 2014; Speer et al., 2017; Sap et al., 2019). A carefully designed relation set is indispensable for many fundamental tasks such as knowledge graph construction. However, it is still unclear whether we need to explicitly represent relations if the final goal is to improve downstream tasks (e.g., machine reading comprehension) that do not directly depend on the reliability of relations in triples from other sources. Once we decide not to name relations, one natural question is whether we could implicitly represent relations between two phrases. We suggest that adding context in which the phrases occur may be useful as such a context constrains the possible relations between phrases without intervening in the relations explicitly (Brézillon et al., 1998). Hereafter, we call a triple that contains a phrase pair and its associated context as a piece of **contextualized knowledge**.

Besides verbal information that is written or spoken, it is well accepted that nonverbal information is also essential for face-to-face communication (Jones and LeBaron, 2002). We regard related verbal and nonverbal information as the phrase pair; we treat the context in which the verbal-nonverbal pair occurs as the context. Such a triple can be regarded as a piece of commonsense knowledge as verbal and nonverbal information function together in communications, and this kind of knowledge is assumed to be known by most people without being formally taught. For example, as shown in Table 1, the pause in *"I'm going......to his house."* is related

---

† Work was conducted when K. S. was an intern at the Tencent AI Lab, Bellevue, WA. Equal contribution.

scene	
□	Interior. Runaway office. Day.
Andy:	I tried to ask her, but...
Emily:	You never ask Miranda. Anything. (**sighs**) All right, I'll take care of the other stuff. You go to Calvin Klein.
Andy:	Me?
Emily:	I'm sorry. Do you have a prior commitment? Is there some hideous pants convention?
Andy:	So I just, what, go down to the Calvin Klein store and ask them...
◇	**Emily rolls her eyes so hard they almost eject from her head.**
Emily:	You're not going to the store.
Andy:	Of course not. I'm going...(**thinking**)...to his house.
Emily (**oh god**):	You are catching on quickly. We always send assistants to a designer's home on their very first day. You're going to his showroom. I'll give you the address.
Andy:	Sorry. Got it. What's the nearest subway stop?
Emily:	Good God. You do not. Under any circumstances. Take public transportation.
Andy:	I don't?

type	nonverbal	verbal
B_c	oh god	Emily: You are catching on [...] I'll give you the address.
I	sighs	Emily: You never ask Miranda. Anything. All right [...] Klein.
I	thinking	Andy: Of course not. I'm going......to his house.
O	Emily rolls her eyes so hard they almost eject from her head.	Andy: So I just, what, go down to the Calvin Klein store and ask them...

Table 1: A sample scene in a script and examples of verbal-nonverbal pairs extracted from this scene (all translated into English; [...]: words omitted; □: scene heading; ◇: action line). The scene is regarded as the context of all the verbal-nonverbal pairs.

to *"thinking"*, the internal state of the speaker. We suggest film and television show scripts are good source corpora for extracting contextualized commonsense knowledge as they contain rich strongly interrelated verbal (e.g., utterances of speakers) and nonverbal information (e.g., body movements, vocal tones, or facial expressions of speakers), which is originally conveyed in different modalities within a short time period and can be easily separated from the scripts. Furthermore, a script usually contains multiple scenes, and the entire text of the scene from which the verbal-nonverbal pair is extracted can serve as the context. According to the relative position of a verbal-nonverbal pair in a scene, we use lexical patterns to extract four types of contextualized knowledge (Section 2).

To use contextualized knowledge to improve MRC, we randomly select nonverbal messages from the same script to convert each piece of knowledge into a weakly-labeled MRC instance (Section 3). We propose a two-stage fine-tuning strategy to use the weakly-labeled MRC data: first, we train a model on the combination of the weakly-labeled data and the target MRC data that is human-annotated but relatively small-scale, and then, we fine-tune the resulting model on the target data alone (Section 4). We observe that training over the combination of all the data based on all types of contextualized knowledge does not lead to notice-

able gains compared to using one type of knowledge. Therefore, we further use a teacher-student paradigm with multiple teacher models trained with different types of knowledge (Section 5).

We evaluate our method on a multiple-choice MRC dataset C^3 (Sun et al., 2020) in which most questions require prior knowledge such as commonsense knowledge besides the given contents. Experimental results demonstrate that our method leads to a 4.3% improvement in accuracy over a state-of-the-art baseline (Sun et al., 2020; Cui et al., 2020). We also seek to transfer the knowledge to a different task by adapting the resulting student MRC model, which yields a 2.9% improvement in F1 over a baseline on a dialogue-based relation extraction dataset DialogRE (Yu et al., 2020).

The main contributions are as follows: (**i**) we suggest that scripts can be a good resource for extracting contextualized commonsense knowledge, and our empirical results demonstrate the usefulness of contextualized knowledge for MRC tasks that require commonsense knowledge and the feasibility of implicitly representing relations by situating structured knowledge in a context; (**ii**) we propose a simple yet effective two-stage fine-tuning strategy to use large-scale weakly-labeled data; and (**iii**) we further show the effectiveness of a teacher-student paradigm to inject multiple types of contextualized knowledge into a single model.

2 Contextualized Knowledge Extraction

Both verbal information that is written or spoken and nonverbal information (e.g., body movements and facial expressions) are essential for face-to-face communication (Jones and LeBaron, 2002; Calero, 2005). We propose to use interrelated verbal and nonverbal information as phrases in the traditional form of commonsense knowledge representation (Speer et al., 2017).

We regard the interrelationship between such a verbal-nonverbal pair as a kind of commonsense knowledge because they function together in communications, and such knowledge is assumed to be known by most people without being formally told just as the definition of commonsense knowledge. We now introduce how to extract verbal-nonverbal pairs and extract the context in which it occurs. Formally we call a triple (v, c, n) as a piece of **contextualized knowledge**, containing a pair of related verbal information v and nonverbal information n, as well as the associated context c. We choose to extract contextualized knowledge from film and television show scripts[1] as rich verbal and nonverbal messages frequently co-occur in scripts, and they can be easily separated. Scenes in a script are separated by blank lines. According to the relative position of verbal and nonverbal information, we extract four types of contextualized knowledge $(B_c, B_n, I, \text{and } O)$ as follows.

- Beginning: the nonverbal information n appears after a speaker name and before the speaker's utterance. We regard the speaker name and the corresponding utterance as v.

    ○ Clean (B_c): We only extract nonverbal information n within parentheses.
    ○ Noisy (B_n): The first span of a turn, followed by a colon, can also contain both a speaker name and nonverbal information about this speaker. It usually happens when a script is written without strictly following a standard screenplay format. We remove the phrase that is a potential speaker name from the span and regard the remaining text in the span as n. We roughly regard a phrase as a speaker name if it appears in the first span of other turns in the same scene.

- Inside (I): We only extract nonverbal information n enclosed in parentheses, which appears within an utterance. All the information in the same turn except n is treated as v.

- Outside (O): Here n is an action line that mainly describes what can be seen or heard by the audience, marked by ◇ in Table 1. We regard the turn (if it exists) before the action line as its corresponding v.

We do not extract phrases in parentheses or action lines as nonverbal information if they are terminologies for script writing such as *"O.S."*, *"CONT'D"*, *"beat"*, *"jump cut"*, and *"fade in"*.[2] All types of contextualized knowledge extracted from a scene share the same context c, i.e., the scene itself. We do not exploit the scene heading mostly about when and where a scene takes place (marked by □ in Table 1), as it is intentionally designed to cover the content of the whole scene, which is already used as context.

3 Instance Generation

As most current MRC tasks requiring commonsense knowledge are usually in a multiple-choice form, we mainly discuss how to convert the extracted triples into multiple-choice instances and leave its extension to other types (e.g., extractive or abstractive) of MRC tasks for future research.

We generate instances for each type of contextualized knowledge. For each triple (v, c, n), we remove n from context c, and we regard the remaining content as the reference document, verbal information v as the question, and the nonverbal information n as the correct answer option. To generate distractors (i.e., wrong answer options), we randomly select N items from all the unique nonverbal information in other triples, which belong to the same type of contextualized knowledge and are extracted from the same script as (v, c, n). Note that we only generate one instance based on each triple, while it is easy to generate more instances by changing distractors.

4 Two-Stage Fine-Tuning

As mentioned previously, we aim to use the constructed weakly-labeled data to improve a downstream MRC task (C^3 in this paper). Given weakly-labeled data generated based on **one** type of contextualized knowledge (e.g., B_c or I) extracted from

---

[1]As it is difficult to verify whether a text is written before a presentation (i.e., script) or during/after a presentation (i.e., transcript), we use *scripts* throughout this paper.

[2]We will release the stop word list along with the code.
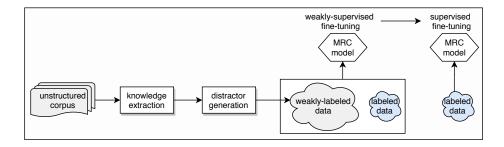
Figure 1: Two-stage fine-tuning framework overview (one type of contextualized knowledge is involved).

scripts, we first use the weakly-labeled data in conjunction with the training set of C^3 as the training data to train the model and then fine-tune the resulting model on C^3 as illustrated in Figure 1. We do not adjust the ratio of clean data to weakly-labeled data observed during training as previous joint training work on other tasks such as machine translation (Edunov et al., 2018).

Another way is to perform separate training: we first train the model on the weakly-labeled data and then fine-tune it on C^3. In our preliminary experiment, we observe that joint training leads to better performance, and therefore we apply it in all the experiments. See performance comparisons of joint and separate training in Section 6.4.

5 Teacher-Student Paradigm

As introduced in Section 3, we have **multiple** sets of weakly-labeled data, each corresponding to one type of contextualized knowledge. We observe that simply combining all the data, either in joint training or separate training, does not lead to noticeable gains compared to using one type of contextualized knowledge. Inspired by the previous work (You et al., 2019) that trains a student automatic speech recognition model with multiple teacher models, and each teacher model is trained on a domain-specific subset with a unique speaking style, we employ a teacher-student paradigm to inject multiple types of contextualized knowledge into a single student machine reader.

Let V denote a set of labeled instances, $W_1, \ldots, W_\ell$ denote ℓ sets of weakly-labeled instances, and $W = \bigcup_{1 \le i \le \ell} W_i$. For each instance t, we let m_t denote its total number of answer options, and $\boldsymbol{h}^{(t)}$ be a hard label vector (one-hot) such that $\boldsymbol{h}_j^{(t)} = 1$ if the j-th option is labeled as correct. We train ℓ teacher models, denoted by $\mathcal{T}_1, \ldots, \mathcal{T}_\ell$, and optimize $\mathcal{T}_i$ by minimizing $\sum_{t \in V \cup W_i} L_1(t, \theta_{\mathcal{T}_i})$.

L_1 is defined as

$$L_1(t, \theta) = - \sum_{1 \le k \le m_t} \boldsymbol{h}_k^{(t)} \log p_\theta(k \,|\, t),$$

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

We define soft label vector $\boldsymbol{s}^{(t)}$ such that

$$\boldsymbol{s}_k^{(t)} = \begin{cases} \lambda \, \boldsymbol{h}_k^{(t)} + (1-\lambda) \sum_{1 \le j \le \ell} \frac{1}{\ell} p_{\theta_{\mathcal{T}_j}}(k \,|\, t) & t \in V \\ \lambda \, \boldsymbol{h}_k^{(t)} + (1-\lambda) p_{\theta_{\mathcal{T}_i}}(k \,|\, t) & t \in W_i \end{cases},$$

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

We then train a student model, denoted by $\mathcal{S}$, in a two-stage fashion. In stage one (i.e., weakly-supervised fine-tuning), we optimize $\mathcal{S}$ by minimizing $\sum_{t \in V \cup W} L_2(t, \theta_{\mathcal{S}})$, where L_2 is defined as

$$L_2(t, \theta) = - \sum_{1 \le k \le m_t} \boldsymbol{s}_k^{(t)} \log p_\theta(k \,|\, t).$$

In stage two (i.e., supervised fine-tuning), we further fine-tune the resulting $\mathcal{S}$ after stage one by minimizing $\sum_{t \in V} L_2(t, \theta_{\mathcal{S}})$. See Figure 2 for an overview of the paradigm.

6 Experiment

6.1 Data

We collect 8,166 scripts in Chinese, and most of them are intended for films and television shows.[3] After segmentation and filtering, we obtain 199,280 scenes, each of which contains at least one piece of contextualized knowledge defined in Section 2. We generate four sets of weakly-labeled data based on the scenes. For comparison, we also use existing human-annotated triples about commonsense
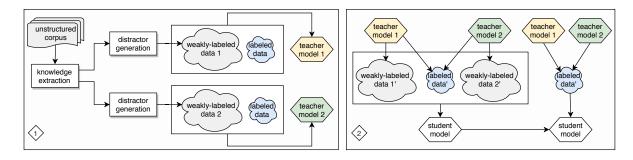
---

[3]https://www.1bianju.com.

Figure 2: Teacher-student paradigm overview (multiple types of contextualized knowledge are involved). To save space, we only show the case that involves two types of contextualized knowledge.

knowledge in the Chinese version of Concept-Net (Speer et al., 2017). We set the number of distractors N (Section 3) to five when we convert structured triples into MRC instances.

For evaluation, we use C^3, a free-form multiple-choice MRC data for Chinese collected from Chinese-as-a-second-language exams (Sun et al., 2020). About 86.8% of questions in C^3 involve prior knowledge (i.e., linguistic, domain-specific, and commonsense knowledge) not provided in the given texts, and all instances are carefully designed by experts such as second-language teachers. Each instance consists of a document, a question, and multiple answer options; only one answer option is correct. See Table 2 for data statistics.

While we focus on scripts and datasets in Chinese in this study, our extraction and training methods are not limited to a particular language.

data	type of construction	# of instances
C^3	human-annotated	19,577
ConceptNet	human-annotated	737,534
B_c	weakly-labeled	105,622
B_n	weakly-labeled	198,053
I	weakly-labeled	204,750
O	weakly-labeled	192,391
B_c+ B_n+ I + O	weakly-labeled	700,816

Table 2: Data statistics.

6.2 Implementation Details

In our experiments, we follow Sun et al. (2020) for the model architecture consisting of a pre-trained language model and a classification layer on top of the model. We use RoBERTa-wwm-ext-large (Cui et al., 2020) as the pre-trained language model, which achieves state-of-the-art performance on C^3 and many other natural language understanding tasks in Chinese (Xu et al., 2020). We leave the ex-

ploration of more pre-trained language models for future work. When the input sequence length exceeds the limit, we repeatedly discard the last turn in the context, or the first turn if the last turn includes the extracted verbal information. We train a model for one epoch during the weakly-supervised fine-tuning stage and eight epochs during the supervised fine-tuning stage. We set λ (defined in Section 5) to 0.5 in all experiments based on the rationale that we can make best use of the soft labels while at the same time making sure $\arg\max_k \boldsymbol{s}_k^{(t)}$ is always the index of the correct answer option for instance t. Carefully tuning λ on the development set may lead to further improvements, which is not the primary focus of this paper.

6.3 Main Results and Discussions

Table 3 reports the main results. The baseline accuracy (73.4% $\{0\}$) is slightly lower than previously reported using the same language model[4] as we report the average accuracy over five runs with different random seeds for all our supervised fine-tuning results. For easy reference, we indicate the index for each result in curly brackets in the following discussion. Obviously, the performance of a model after the first fine-tuning stage over the combination of the C^3 dataset and much larger weakly-labeled data is worse (e.g., 71.7% $\{1\}$) than baseline performance ($\{0\}$). Further fine-tuning the resulting model on the C^3 dataset consistently leads to improvements (e.g., 74.0% $\{2\}$ and 74.5% $\{4\}$) over the baseline $\{0\}$, demonstrating the effectiveness of the **two-stage** fine-tuning strategy for using large-scale weakly-labeled data. We will discuss the critical role of the target task's data (i.e., C^3) in the weakly-supervised fine-tuning stage in the next subsection. Following this strategy, **each** of the weakly-labeled data based on one

---

[4]https://github.com/CLUEbenchmark/CLUE.

index	weakly-supervised fine-tuning		supervised fine-tuning		dev	test
	data	teacher-student	data	teacher-student		
0	–	–	C^3	–	73.9	73.4
1	$C^3 + B_c$	–	–	–	71.1	71.7
2	$C^3 + B_c$	–	C^3	–	74.5	74.0
3	$C^3 + B_n$	–	–	–	71.3	72.0
4	$C^3 + B_n$	–	C^3	–	74.6	74.5
5	$C^3 + I$	–	–	–	73.5	72.8
6	$C^3 + I$	–	C^3	–	**75.6**	**74.9**
7	$C^3 + O$	–	–	–	72.4	72.7
8	$C^3 + O$	–	C^3	–	75.4	74.9
9	$C^3 + B_c + B_n + I + O$	–	–	–	71.6	71.0
10	$C^3 + B_c + B_n + I + O$	–	C^3	–	75.6	75.2
11	$C^3 + B_c + B_n + I + O$	✓	C^3	–	76.5	76.4
12	$C^3 + B_c + B_n + I + O$	✓	C^3	✓	**77.4**	**77.7**

Table 3: Average accuracy (%) on the development and test sets of the C^3 dataset.

type of contextualized knowledge can boost the final performance ({2, 4, 6, 8}); the magnitude of accuracy improvement is 1.2% on average.

When we combine all the weakly-labeled data in the first fine-tuning stage, the performance gain after the second round of fine-tuning (75.2% {10}) is not as impressive as expected, given the best performance achieved by only using one set (74.9% {6}). As a comparison, our **teacher-student paradigm** that trains multiple teacher models with different types of weakly-labeled data leads to up to 3.7% improvement in accuracy ({12} vs. {2, 4, 6, 8}). The advantage is reduced but still exists even when we use the original hard labels instead of soft labels in the second fine-tuning stage (76.4% {11}).

6.4 Ablation Studies and Discussions

We have shown that our proposed teacher-student paradigm helps inject multiple types of knowledge into the baseline. We conduct ablation studies to examine critical factors. We first remove the context (i.e., scene) from each instance in the weakly-labeled data and leave it empty. All other aspects of this baseline remain the same as {12} in Table 3. We also remove the C^3 dataset from the weakly-supervised fine-tuning stage when we train teacher and student models (Figure 2) and only use C^3 during the supervised fine-tuning stage. We observe that accuracy decreases in both conditions (Table 4), demonstrating the usefulness of contexts in contextualized knowledge for improving machine reading comprehension and the importance of involving the human-annotated data of the target task, although small-scale, in the weakly-supervised fine-tuning stage.

method	dev	test
{12} in Table 3	77.4	77.7
{12} w/o context in weakly-labeled data	76.8	76.6
{12} w/o using C^3 in the 1st FT	76.6	76.2

Table 4: Ablation results on the development and test sets of the C^3 dataset (FT: fine-tuning).

category	{0}	{12}	Δ
Matching	90.0	**94.7**	4.7
Prior Knowledge	69.5	**75.3**	5.8
⋄ Linguistic	73.8	**77.8**	4.0
⋄ Commonsense	68.0	**74.4**	6.4
⋄ Domain-specific⋆	13.3	**20.0**	6.7

Table 5: Average accuracy (%) on the annotated development set of C^3 per question category (⋆: the domain-specific category only contains three instances).

As we may require one or multiple types of prior knowledge to answer a question, we study the impacts of the contextualized knowledge on different types of questions based on the annotated subset (300 instances) released along with the dataset. As shown in Table 5, our method generally improves performance on all types of questions, especially those that require commonsense knowledge.

Considering the similarity of B_c and B_n in the relative position of verbal and nonverbal information in a scene, we also experiment by merging B_c and B_n into a single set and then training three teacher models instead of four used for training the student model. Results show that it achieves a similar accuracy (77.5%) to the four-teacher setting ({12} in Table 3). For further improvement, it may be a promising direction to train more teachers

notes	weakly-labeled data				dev	test
	structured knowledge	document	question	answer		
{0} in Table 3	–	–	–	–	73.9	73.4
{10} in Table 3	contextualized knowledge	scene	verbal	nonverbal	75.6	75.2
{10} w/o context	contextualized knowledge	empty	verbal	nonverbal	74.9	74.2
i	ConceptNet	empty	subject	object	74.0	72.7
ii	ConceptNet	relation type	subject	object	74.6	74.1

Table 6: Average accuracy (%) on the development and test sets of the C^3 dataset using weakly-labeled data constructed based on contextualized knowledge or ConceptNet.

with diverse types or forms of external knowledge.

6.5 A Comparison Between Contextualized Knowledge and ConceptNet

Most of the existing commonsense knowledge graphs are in English. Therefore, we only compare contextualized knowledge with the Chinese version of a human-annotated commonsense knowledge graph ConceptNet. Each triple in ConceptNet is represented as (subject, relation type, object) (e.g., (*"drink water"*, CAUSES, *"not thirsty"*)). We experiment with three types of input sequences when we convert triples into MRC instances: (i) leave the document empty in each instance and (ii) use the relation type as the document. We randomly select phrases in ConceptNet other than the phrases in each triple as distractors.

For a fair comparison, we compare (ii) with baseline {10} in Table 3 as it follows the same two-stage fine-tuning without using the teacher-student paradigm. To compare with (i), we run an ablation test of {10} by removing contexts from weakly-labeled MRC instances. The amounts of weakly-labeled instances based on contextualized knowledge and ConceptNet are similar (Table 2). The results in Table 6 reveal that under the two-stage fine-tuning framework, introducing Concept-Net yields up to 0.7% in accuracy, but using contextualized knowledge gives a bigger gain of 1.8% in accuracy. Furthermore, removing contexts from weakly-labeled instances hurts performance, consistent with our observation in Section 6.4.

We do not dismiss the construction and use of commonsense knowledge with a well-defined schema and admit that the form of contextualized knowledge representation is not concise enough for easy alignment with existing commonsense knowledge graphs or knowledge graph completion. However, we argue that contexts can tacitly state the relation between phrases, and this kind of commonsense knowledge is helpful for MRC.

6.6 Transferring to Relation Extraction

Seeking to transfer knowledge to other tasks, we take a relation extraction task DialogRE (Yu et al., 2020) as a case study. The task aims to predict relations between an argument pair based on a given dialogue. We replace the classification layer of an MRC model with a multi-class multi-label classification layer following the baseline released by Yu et al. (2020) and fine-tune the whole architecture on the Chinese version of the DialogRE dataset.

We compare the performance of methods that use different weights for parameter initialization except for the classification layer, which is randomly initialized. As shown in Table 7, we achieve an improvement of 2.9% in F1 and 3.1% in $F1_c$ by adapting our best-performing machine reading comprehension model. The metric $F1_c$ is used to encourage a model to identify relations between arguments as early as possible rather than after reading the whole dialogue. Introducing C^3 alone also allows us to achieve a slight gain over the baseline. It might be interesting to study the relevance between document/dialogue-based relation extraction and machine reading comprehension to boost the performance of the two types of tasks.

parameter initialization	dev		test	
	F1	$F1_c$	F1	$F1_c$
RoBERTa-wwm-ext-large	64.9	60.3	64.4	59.2
{0} in Table 3	66.4	61.6	65.0	60.3
{12} in Table 3	**67.1**	**62.9**	**67.3**	**62.3**

Table 7: Average F1 (%) and $F1_c$ (%) on the DialogRE dataset.

7 Related Work

7.1 Contextualized Knowledge

We mainly discuss the external contextualized knowledge that is not directly relevant with a target task as retrieving relevant pieces of evidence from

an external source for instances of a target task is not the focus of this paper. A common solution to obtain external contextualized knowledge is to utilize existing knowledge bases via distant supervision. For example, Ye et al. (2019) align triples in ConceptNet (Speer et al., 2017) with sentences from Wikipedia. We extract contextualized knowledge from scripts, where contexts (i.e., scenes) are naturally aligned with verbal-nonverbal pairs to avoid noise from distant supervision.

Our work is also related to commonsense knowledge extraction, which relies on human-annotated triples (Xu et al., 2018; Bosselut et al., 2019), high-precision syntactic or semantic patterns (Zhang et al., 2020; Zhou et al., 2020) specific to each relation, or existing lexical databases (Tandon et al., 2014, 2015). In comparison, we skip the step of offering a name of the relation between two phrases and focus on extracting structured knowledge in its context. Our language-independent knowledge extraction does not require any training data and does not rely on a high-quality semantic lexicon or a syntactic parser, which is usually unavailable for many non-English languages.

7.2 Weak Supervision for Machine Reading Comprehension

As it is expensive and time-consuming to crowd-source or collect a large-scale, high-quality dataset, weak supervision has received much attention throughout the history of machine reading comprehension. Various forms of weak supervision are studied, mostly based on existing resources such as pre-trained semantic/syntactic parsers (Smith et al., 2015; Wang et al., 2015; Liu et al., 2017) or natural language inference systems (Pujari and Goldwasser, 2019; Wang et al., 2019), knowledge bases (Wang et al., 2018b; Wang and Jiang, 2019; Yang et al., 2019), and linguistic lexicons (Sun et al., 2019b). Compared to previous work, we focus on generating large-scale weakly-labeled data using the contextualized knowledge automatically extracted from unstructured corpora.

7.3 Semi-Supervised Learning for Machine Reading Comprehension

Previous semi-supervised methods that leverage internal or external unlabeled texts usually generate question and answer based on the content of the same sentence (Yang et al., 2017; Wang et al., 2018a; Dhingra et al., 2018). Besides the unlabeled texts, previous studies (Yuan et al., 2017; Yu et al., 2018; Zhang and Bansal, 2019; Zhu et al., 2019; Dong et al., 2019; Alberti et al., 2019; Asai and Hajishirzi, 2020) also heavily rely on the labeled instances of the target machine reading comprehension task for data augmentation. In comparison, we focus on generating non-extractive instances without using any task-specific patterns or labeled data, aiming to improve machine reading comprehension tasks that require substantial prior knowledge such as commonsense knowledge.

Another line of work develops unsupervised approaches (Lewis et al., 2019; Li et al., 2020; Fabbri et al., 2020) for extractive machine reading comprehension tasks. However, there is still a large performance gap between unsupervised and state-of-the-art supervised methods.

7.4 Knowledge Integration

Our teacher-student paradigm for knowledge integration is most related to multi-domain teacher-student training for automatic speech recognition (You et al., 2019) and machine translation (Wang et al., 2020). Instead of clean domain-specific human-labeled data, each of our teacher models is trained with weakly-labeled data. Due to the introduction of large amounts of weakly-labeled data, the data of the target machine reading comprehension task (with hard or soft labels) is used during all the fine-tuning stages of both teacher and student models.

8 Conclusions

In this paper, we aim to extract contextualized commonsense knowledge to improve machine reading comprehension. We propose to situate structured knowledge in a context to implicitly represent the relationship between phrases, instead of relying on a pre-defined set of relations. We extract contextualized knowledge from film and television show scripts as interrelated verbal and nonverbal messages frequently co-occur in scripts. We propose a two-stage fine-tuning strategy to use the large-scale weakly-labeled data and employ a teacher-student paradigm to inject multiple types of contextualized knowledge into a single student model. Experimental results demonstrate that our method outperforms a state-of-the-art baseline by a 4.3% improvement in accuracy on the multiple-choice machine reading comprehension dataset C^3, wherein most of the questions require unstated prior knowledge, especially commonsense knowledge.

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.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the ACL*, pages 4762–4779.

Patrick Brézillon, J-Ch Pomerol, and Ilham Saker. 1998. Contextual and contextualized knowledge: An application in subway control. *International Journal of Human-Computer Studies*, 48(3):357–373.

Henry H Calero. 2005. *The power of nonverbal communication: How you act is more important than what you say.* Silver Lake Publishing.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pretrained models for chinese natural language processing. *arXiv preprint*, cs.CL/2004.13922v1.

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

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the EMNLP*, pages 489–500.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the ACL*, pages 4508–4513.

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

Stanley E Jones and Curtis D LeBaron. 2002. Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of communication*, 52(3):499–521.

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.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the ACL*, pages 4896–4910.

Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. Harvesting and refining question-answer pairs for unsupervised QA. In *Proceedings of the ACL*, pages 6719–6728.

Rui Liu, Junjie Hu, Wei Wei, Zi Yang, and Eric Nyberg. 2017. Structural embedding of syntactic trees for machine comprehension. In *Proceedings of the EMNLP*, pages 815–824.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. In *Proceedings of the NAACL-HLT*, pages 839–849.

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.

Rajkumar Pujari and Dan Goldwasser. 2019. Using natural language relations between answer choices for machine comprehension. In *Proceedings of the NAACL-HLT*, pages 4010–4015.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI*, pages 3027–3035.

Ellery Smith, Nicola Greco, Matko Bošnjak, and Andreas Vlachos. 2015. A strong lexical matching method for the machine comprehension test. In *Proceedings of the EMNLP*, pages 1693–1698.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI*, pages 4444–4451.

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. Probing prior knowledge needed in challenging chinese machine reading comprehension. *arXiv preprint*, cs.CL/1904.09679v2.

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

Niket Tandon, Gerard De Melo, Abir De, and Gerhard Weikum. 2015. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of the CIKM*, pages 223–232.

Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the WSDM*, pages 523–532.

Chao Wang and Hui Jiang. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *Proceedings of the ACL*, pages 2263–2272.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the ACL*, pages 700–706.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the CoNLL*, pages 696–707.

Liang Wang, Sujian Li, Wei Zhao, Kewei Shen, Meng Sun, Ruoyu Jia, and Jingming Liu. 2018a. Multi-perspective context aggregation for semi-supervised cloze-style reading comprehension. In *Proceedings of the COLING*, pages 857–867.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018b. Yuanfudao at SemEval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of the SemEval*, pages 758–762.

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

Frank F. Xu, Bill Yuchen Lin, and Kenny Zhu. 2018. Automatic extraction of commonsense LocatedNear knowledge. In *Proceedings of the ACL*, pages 96–101.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, and Zhenzhong Lan. 2020. CLUE: A chinese language understanding evaluation benchmark. *arXiv preprint*, cs.CL/2004.05986v2.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the ACL*, pages 2346–2357.

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.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint*, cs.CL/1908.06725v5.

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*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the ACL*, pages 4927–4940.

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.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the IJCAI*.

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.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the ACL*, pages 7579–7589.

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.