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

To perform well on a machine reading comprehension (MRC) task, machine readers usually require commonsense knowledge that is not explicitly mentioned in the given documents. This paper aims to extract a new kind of structured knowledge from scripts and use it to improve MRC. We focus on scripts as they contain rich verbal and nonverbal messages, and two relevant messages originally conveyed by different modalities during a short time period may serve as arguments of a piece of commonsense knowledge as they function together in daily communications. To save human efforts to name relations, we propose to represent relations implicitly by situating such an argument pair in a context and call it contextualized knowledge.

To use the extracted knowledge to improve MRC, we compare several fine-tuning strategies to use the weakly-labeled MRC data constructed based on contextualized knowledge and further design a teacher-student paradigm with multiple teachers to facilitate the transfer of knowledge in weakly-labeled MRC data. Experimental results show that our paradigm outperforms other methods that use weakly-labeled data and improves a state-of-the-art baseline by 4.3% in accuracy on a Chinese multiple-choice MRC dataset C^3, wherein most of the questions require unstated prior knowledge. We also seek to transfer the knowledge to other tasks by simply adapting the resulting student reader, yielding a 2.9% improvement in F1 on a relation extraction dataset DialogRE, demonstrating the potential usefulness of the knowledge for non-MRC tasks that require document comprehension.

1 Introduction

Given a question and a document, machine reading comprehension (MRC) tasks usually aim to select the correct answer option(s) from all options (Richardson et al., 2013; Lai et al., 2017) or extract a span as the answer from the document (Hermann et al., 2015; Rajpurkar et al., 2016). For many MRC tasks, which are mostly in multiple-choice formats, machine readers require extensive general world knowledge unstated in the given documents to perform well on tasks. Recent studies have showed the usefulness of other human-annotated MRC datasets (Chung et al., 2018) or data from relevant tasks such as natural language inference (Yin et al., 2021). There is also a trend towards taking advantage of existing crowdsourced general world knowledge graphs such as ConceptNet (Speer et al., 2017) or automatically constructed graphs (Zhang et al., 2020) to improve MRC tasks (Mostafazadeh et al., 2016; Ostermann et al., 2018; Sun et al., 2019a; Huang et al., 2019).

However, most progress has been limited to English, and it is expensive, time-consuming, laborious, and error-prone to construct clean, large-scale datasets of MRC-related tasks or general world knowledge graphs, which are not always available even for high-resource languages such as Chinese. This paper aims to extract a new kind of structured general world knowledge from external unstructured Chinese corpora, which is seldom studied, and investigate the use of the knowledge-based augmented data to improve MRC.

Typically, each piece of general world 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). However, naming relations often requires substantial human efforts (e.g., guidelines, annotations, or relation-specific patterns), and it is also unclear whether we need to explicitly represent relations if the final goal is to improve downstream tasks that do not directly depend on the reliability of relations in triples from

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

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8736 - 8747

other sources. Once we have decided 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** for convenience.

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 triples contain rich 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, just as commonsense knowledge. For example, as shown in Table 1, the pause in *"I'm going.....to his house."* is related to *"thinking"*, the internal state of the speaker. We suggest film and television show scripts are good source corpora for extracting contextualized 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 extracted 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 create four types of simple lexical patterns to extract contextualized verbal-nonverbal knowledge (Section 2).

To use contextualized knowledge to improve MRC, we randomly select nonverbal messages from the same script as distractors to convert each piece of knowledge into a weakly-labeled MRC instance (Section 3). We explore different two-stage fine-tuning strategies to use the weakly-labeled MRC data: for example, we first train a model on the combination of the weakly-labeled data and the target MRC data that is human-annotated but small-scale, and then, we fine-tune the resulting model on the target data (Section 4). However, we observe that increasing the amounts of weakly-labeled data does not lead to noticeable gains than using one subset of the data. Inspired by teacher-student studies that facilitate the transfer of **clean** knowledge (Li et al., 2014; Hinton et al., 2015; You et al., 2019), we further design a teacher-student paradigm with multiple teachers, each trained with a subset of the hard-labeled, **weakly-labeled** data, and we feed soft-labeled, weakly-labeled/clean data to a student in a two-stage fashion, similar to the two-stage fine-tuning strategy mentioned earlier (Section 5).

We evaluate our method on C^3, a representative multiple-choice general-domain MRC dataset for Chinese wherein most questions require commonsense knowledge beyond the given contents (Sun et al., 2020). To date, C^3 is the only public Chinese dataset of this task. Experimental results show that our multi-teacher paradigm leads to $+4.3\%$ in accuracy over a state-of-the-art baseline (Cui et al., 2020) and also outperforms other methods such as pre-training upon all scripts and hard-label fine-tuning that uses the same structured knowledge. Also, under the same framework, constructing weakly-labeled MRC data based on verbal-nonverbal knowledge leads to bigger gains compared to other strong data augmentation strategies (and some are even based on human-annotated commonsense knowledge graphs (Speer et al., 2017)). Finally, we seek to transfer the knowledge to help other tasks by adapting the final student MRC model, yielding $+2.9\%$ in F1 on the Chinese set of a document-level relation extraction dataset DialogRE (Yu et al., 2020) over competitive baselines.

The main contributions are as follows: (**i**) we suggest that scripts can be a good resource for extracting rich contextualized verbal-nonverbal knowledge with minimal supervision; (**ii**) we explore and compare several fine-tuning strategies to use weakly-labeled data based on the extracted knowledge and further propose a teacher-student paradigm with multiple teachers to better learn from large-scale weakly-labeled data; and (**iii**) our empirical results demonstrate the effectiveness of the multi-teacher paradigm and the usefulness of contextualized verbal-nonverbal knowledge for MRC tasks that require commonsense knowledge.

2 Contextualized Knowledge Extraction

Understanding the interactions between verbal and nonverbal information in communications requires prior knowledge, and this knowledge is assumed

Scene 1		
□	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?	
pattern 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 extracted verbal-nonverbal pairs from this scene (all translated into English; [...]: words omitted; □: scene heading; ◇: action line). The scene is the context of all pairs.

to be known by most people without being taught, just as commonsense knowledge. We propose to use interrelated verbal and nonverbal information as phrases in the classical triple-style knowledge representation and situate them in a context. 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, and the associated context c. We choose to extract contextualized knowledge from film and television show scripts as plentiful verbal and nonverbal messages frequently co-occur in scripts, and they can be easily separated. Scenes in a script are delimited by blank lines. Based on the relative position of verbal and nonverbal information (intra-turn or cross-turn), we extract four subsets of contextualized knowledge (B_c, B_n, I, and O):

- 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. For example, from *"Xiaocong Le took the cup of hot water: 'Thank you!' "*, *"Xiaocong Le"* will be recognized as the speaker name; *"took the cup of hot water"* will be treated 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."*, *"V.O."* *"CONT'D"*, *"beat"*, *"jump cut"*, and *"fade in"*.[1] All contextualized knowledge triples extracted from a scene share the same context c, i.e., the scene itself. We do not exploit scene headings (e.g., when and where a scene takes place) (marked by □ in Table 1), as they are intentionally designed to cover the content of a whole scene, which is already used as context. We leave other types of patterns (e.g., using the turn after the action line to construct O) for future studies. Due to space limitations, we list more extracted contextualized verbal-nonverbal knowledge triples in Appendix A.3 (Table 11).

[1] We will release the stop word list along with the code at https://github.com/nlpdata/script.

3 MRC Instance Generation

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 one instance for each piece 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 messages in other triples, which are extracted using the same type of patterns from the same script as (v, c, n). Note that although we only generate one instance based on each triple, it is easy to generate more instances by changing distractors.

4 Two-Stage Fine-Tuning

We aim to use the constructed weakly-labeled data to improve a downstream MRC task. Given weakly-labeled data generated based on contextualized knowledge extracted from scripts, we first use the weakly-labeled data in conjunction with the training set of the target MRC data as the training data to train the model and then fine-tune the resulting model on the target MRC data 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 (or sequential transfer learning (Ruder et al., 2019)): we can first train the model on the weakly-labeled data and then fine-tune it on the target data. We assume that using the small-scale, clean data to guide weakly-supervised training is more helpful, and we compare joint and separate training in Section 6.5.

5 Multi-Teacher Paradigm

In our preliminary experiment, we observe that increasing the amounts of weakly-labeled data does not lead to noticeable gains than using one subset of the weakly-labeled data. Inspired by previous teacher-student frameworks (You et al., 2019; Wang et al., 2020) that train a multi-domain student model with multiple teacher models to help knowledge transfer, and each teacher model is trained on a **clean** domain-specific subset, we extend the multi-teacher idea to let models learn better from large-scale **weakly-labeled** MRC data for improving a target MRC task.

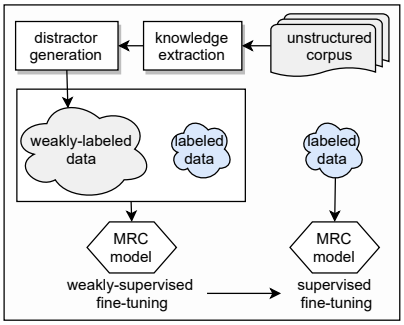

Figure 1: Two-stage fine-tuning framework overview (hard labels are used during both stages).

As introduced in Section 3, we can have multiple subsets of weakly-labeled data generated based on verbal-nonverbal knowledge extracted by different types of patterns ((e.g., B_c or I) in (Section 2)).

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 $\boldsymbol{h}^{(t)}_j = 1$ 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

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

where $p_\theta(k \mid t)$ 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

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

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

We then train a student model, denoted by $\mathcal{S}$, still 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

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

In stage two (i.e., supervised fine-tuning), we further fine-tune the resulting $\mathcal{S}$ after stage one by

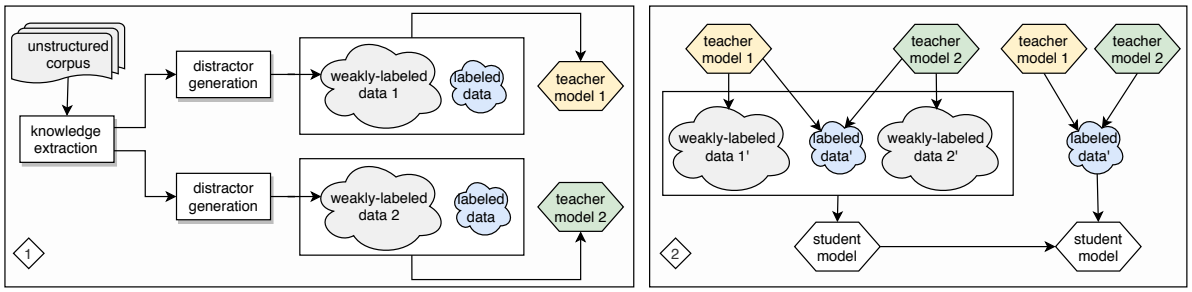

Figure 2: Multi-teacher paradigm overview (multiple subsets of contextualized knowledge are involved). To save space, we only show the case that involves two subsets. Teachers (1 and 2) are trained with hard-labeled data, and one student is trained with soft-labeled data in both stages.

minimizing $\sum_{t \in V} L_2(t, \theta_S)$. See Figure 2 for an overview of the multi-teacher paradigm.

6 Experiment

6.1 Data

data	type of construction	# of instances
C^3	human-annotated	19,577
DialogRE	human-annotated	10,886
ConceptNet	human-annotated	737,534
span-style augmented data	weakly-labeled	894,834
B_c+ B_n+ I + O	weakly-labeled	700,816
B_c	weakly-labeled	105,622
B_n	weakly-labeled	198,053
I	weakly-labeled	204,750
O	weakly-labeled	192,391

Table 2: Data statistics.

We collect 100,46 scripts in Chinese, and most of them are intended for films and television shows.[2] After segmentation and filtering, we obtain 199,280 scenes, each of which contains at least one piece of contextualized knowledge. We generate four subsets of weakly-labeled MRC data based on contextualized knowledge extracted by four types of patterns. For comparisons, we use existing human-annotated triples about commonsense knowledge in the Chinese version of ConceptNet (Speer et al., 2017) as well as span-style augmented multiple-choice MRC data (Sun et al., 2019b) extracted from all scripts. We set the number of distractors N to five for weakly-labeled MRC instances.

For evaluation, we use C^3, so far as we know, the only public multiple-choice MRC data for Chinese with a focus on commonsense knowledge. About 86.8% of questions in C^3 involve prior knowledge (i.e., linguistic, domain-specific, and commonsense knowledge) unstated 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 option is correct. In addition, we use the human-annotated Chinese set of a bilingual relation extraction dataset DialogRE, which also requires document-level understanding to predict relations from 36 possible types between an argument pair.[3] We provide data statistics in Table 2. While we focus on resources (e.g., scripts and datasets) in Chinese, our extraction and training methods are not limited to a particular language.

6.2 Implementation Details

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 exploration of very recent pre-trained language models such as ERNIE 3.0 (Sun et al., 2021) 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. For all our implementations and methods, we train a model for one epoch during the weakly-supervised fine-tuning stage and eight epochs during the supervised fine-tuning stage; other parameters remain unchanged. 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 option for instance t. Carefully tuning λ on the development set may lead to further improvements,

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

[3] https://github.com/nlpdata/dialogre.

index	name	weakly-supervised fine-tuning		supervised fine-tuning		dev	test
		training data	labels	training data	labels		
T0	regular fine-tuning†	–	–	□	hard	73.9 (0.5)	73.4 (0.5)
1	continued MLM pre-training†	scripts	hard	□	hard	72.9 (0.6)	72.7 (0.9)
2	two-stage fine-tuning	□ + B_c+ B_n+ I + O	hard	□	hard	**75.6** (0.2)	**75.2** (0.6)
3	two-stage fine-tuning	□ + B_c	hard	□	hard	74.5 (0.3)	74.0 (0.4)
4	two-stage fine-tuning	□ + B_n	hard	□	hard	74.6 (0.5)	74.5 (0.8)
5	two-stage fine-tuning	□ + I	hard	□	hard	**75.6** (0.4)	**74.9** (0.6)
6	two-stage fine-tuning	□ + O	hard	□	hard	75.4 (0.3)	74.9 (0.1)
7	two-stage fine-tuning	□ + span-style augmented data (Sun et al., 2019b) based on scripts	hard	□	hard	72.6 (0.4)	72.0 (0.6)
8	two-stage fine-tuning	□ + ConceptNet (Speer et al., 2017) (i)	hard	□	hard	74.0 (0.2)	72.7 (0.4))
9	two-stage fine-tuning	□ + ConceptNet (Speer et al., 2017) (ii)	hard	□	hard	74.6 (0.4)	74.1 (0.5)
10	student with teachers (T1-T4)	□ + B_c+ B_n+ I + O	soft	□	soft	**77.4** (0.3)	**77.7** (0.1)
M1	student with teachers (T1-T4)	□ + B_c+ B_n+ I + O	soft	□	hard	76.5 (0.4)	76.4 (0.3)
M2	student with teachers (4×T0)‡	□	soft	□	soft	72.0 (0.7)	71.5 (1.0)
M3	student with teachers (4×T0)‡	□	hard	□	soft	73.4 (0.3)	73.0 (0.5)
M4	student with teachers (4×T0)‡	–	–	□	soft	72.1 (0.7)	71.7 (0.7)
T1	teacher model	□ + B_c	hard	–	–	71.2 (0.1)	71.6 (0.1)
T2	teacher model	□ + B_n	hard	–	–	72.2 (1.2)	72.3 (0.4)
T3	teacher model	□ + I	hard	–	–	72.9 (0.9)	72.5 (0.4)
T4	teacher model	□ + O	hard	–	–	73.1 (0.9)	72.6 (0.2)

Table 3: Average accuracy with standard deviation (%) on the development and test sets of the C^3 dataset (□: C^3; H: hard; S: soft; †: we implement or use the public implementations of previous work (Xu et al., 2020) for T0 and (Xu et al., 2020) for 1; ‡ four models are trained with different random seeds).

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% {T0}) 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.

Two-Stage Fine-Tuning: 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., 71.6% {T1}) than baseline performance ({T0}). Further fine-tuning the resulting model on the C^3 dataset consistently leads to improvements (e.g., 74.0% {3}) over the baseline {T0}, demonstrating the necessity of supervised fine-tuning when large-scale weakly-labeled data is used. In particular, each subset of the weakly-labeled data based on contextualized knowledge extracted by a certain type of patterns can boost the performance ({3, 4, 5, 6}); the magnitude of accuracy improvement is 1.2% on average.

Multi-Teacher Paradigm: When we combine all the weakly-labeled data during weakly-supervised fine-tuning, the performance gain after the supervised fine-tuning stage (75.2% {2}) is not as impressive as expected, given the best performance achieved by only using one subset (74.9% {5}). As a comparison, our teacher-student paradigm with multiple teacher models (i.e., {T1, T2, T3, T4}) trained with different subsets of weakly-labeled data leads to up to 3.7% improvement in accuracy ({10} vs. {3, 4, 5, 6}). 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% {M1}). To examine the contribution of the contextualized knowledge, under the *exact same* multi-teacher paradigm, we train four {T0}-teachers only on the target MRC dataset with different random seeds, and let them teach the student. Such multi-teacher baselines ({M2–M4}) perform much worse than our proposed method {10}, even under-performing the original {T0}.

Pre-Training or Fine-Tuning: We also consider continuing pre-training the pre-trained language model (i.e., RoBERTa-wwm-ext-large) over the plain texts of the scripts using the masked language model (MLM) objective (Devlin et al., 2019; Liu et al., 2019), which might benefit more downstream tasks without requiring task-specific data augmentation. We follow previous work (Cui et al., 2020; Xu et al., 2020) to conduct whole word masking to generate training instances and continue pre-training the model for 200K steps considering the much smaller size of the scripts (0.1B tokens) compared with that of the original pre-training data (5.4B tokens).[5] As shown in {1} in Table 3, con-

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

[5] Here, increasing the number of steps further hurts MRC.

tinued pre-training over the scripts has a slight negative impact on MRC performance. This result is perhaps not too surprising as many studies demonstrate the usefulness of continued pre-training only when the size of domain-specific unlabeled data (e.g., NEWS (Zellers et al., 2020) that contains 6.7B tokens) is sufficient to update model parameters using MLM (Gururangan et al., 2020). Considering the size of scripts, to improve the performance on a general-domain MRC task, we argue that it is more effective to inject the knowledge during the fine-tuning stage, using the same training objective as that of the target MRC task.

6.4 Comparisons with Other Types of Augmented MRC Data

From the perspective of improving a target MRC task, we first compare the extracted contextualized verbal-nonverbal knowledge with **crowdsourced** commonsense knowledge graphs. As most of such graphs are in English, we only compare the knowledge with the crowdsourced Chinese version of ConceptNet. To convert each triple (subject, relation type, object) in ConceptNet into an MRC instance, we use the subject as question and object as answer, and we experiment with two types of documents: (i) leave the document empty 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 {2} in Table 3 using the same two-stage hard-label fine-tuning strategy. The numbers of weakly-labeled instances based on contextualized knowledge and ConceptNet are close (Table 2). The results reveal that under the two-stage fine-tuning framework, introducing ConceptNet yields up to $+0.7\%$ in accuracy, but using contextualized knowledge gives a bigger gain of 1.8% in accuracy. It is worth mentioning that keeping context (i.e., relation types) for ConceptNet-based weakly-labeled instances is helpful for downstream tasks ({9} vs. {8}).

Based on all the scripts, we also implement a **span-style** multiple-choice MRC data augmentation method (Sun et al., 2019b), which randomly concatenates several sentences in a scene as a question, randomly removes spans from the question, and concatenates them to form an answer. Distractors are generated by randomly replacing spans in the answer with randomly selected spans from the same scene. This method ({7}) does not lead accuracy gains over the baseline {T0}, supporting the usefulness of augmenting MRC data based on verbal-nonverbal knowledge from scripts.

6.5 Ablation Studies and Analysis

We have shown that the present multi-teacher paradigm helps learn better from large-scale weakly-labeled data. We further conduct ablation studies to examine critical factors for this paradigm. All other aspects remain the same as {10} in Table 3. We remove the context (i.e., scene) from each instance in the weakly-labeled data and leave it empty. We also experiment with removing C^3 from the weakly-supervised fine-tuning stage when we train teacher and student models and only use hard-labeled C^3 during the final supervised fine-tuning stage. As shown in Table 4, accuracy decreases in both cases. For two-stage hard-label fine-tuning {2}, keeping context is still helpful, while including C^3 during weakly-supervised fine-tuning seems less important. These results show the usefulness of keeping the context of weakly-labeled data and involving the small-scale, human-annotated target MRC data to guide weakly-supervised fine-tuning for the multi-teacher paradigm.

our methods	dev
{10} in Table 3	77.4 (0.3)
{10} removing context from weakly-labeled data	76.8 (0.2)
{10} removing C^3 from the weakly-supervised fine-tuning	76.6 (0.5)
{10} removing weakly-labeled data (i.e., {M3} in Table 3)	73.4 (0.3)
{2} in Table 3	75.6 (0.2)
{2} removing context from weakly-labeled data	74.9 (0.2)
{2} removing C^3 from the weakly-supervised fine-tuning†	75.5 (0.3)

Table 4: Ablation results on the dev set of C^3 (†: same as sequential transfer learning (Ruder et al., 2019)).

It is difficult, however, to infer which pieces of knowledge help the improved MRC instances. As an alternative solution, 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 6, our method generally improves performance on all types of questions, especially those that require commonsense knowledge.

In particular, we see $+10.0\%$ in accuracy or more on questions that require CAUSE-EFFECT, SCENARIO, or PART-WHOLE, three subcategories of commonsense knowledge. For instance, given a conversation, *"Female: Sir, can you drive faster? I'm afraid that I will be late for the exam. Male: No, the speed is already quite fast. Safety is also very*

Scene 2		
Lili Kang:	Hurry up, the Central Hospital! The child just fell. Hurry up!	
◇	The car is going fast.	
Lili Kang:	Hurry up! hurry up!	
Taxi driver:	I can't go any faster, I'll be fined by the police if I go any faster.	
◇	The taxi driver said and turned to look at the child wrapped in the blanket.	
Taxi driver (**surprised**):	This kid's eyes are wide open, and he is in high spirits. Where did he break it?	
	[...]	
pattern type	**nonverbal**	**verbal**
B_c	surprised	Taxi driver: This kid's eyes are wide open, and he is in high spirits. Where did he break it?

Table 5: An taxi-related example of extracted contexutalized knowledge ([...]: words omitted; ◇: action line).

important.", we require scenario knowledge about activities of humans, their corresponding location information, and personal information such as the profession, in order to answer the question about the possible location (*"taxi"*) of the two speakers. It is possible that the **temporal order** between verbal and nonverbal information sometimes indicates causality. Also, the **scenes** used as context in the structured knowledge contain rich scenario knowledge. For example, similar to the scenario in the previous MRC instance, a taxi driver also refuses the passenger's request to drive faster in one of our extracted triples (Table 5). See more improved MRC instances and analysis in Appendix A.1.

category	{T0}	{10}	Δ
Matching	90.0	**94.7**	4.7
Prior Knowledge	69.5	**75.3**	5.8
· · · Linguistic	73.8	**77.8**	4.0
· · · General world knowledge	68.0	**74.4**	6.4
· · · Arithmetic	34.3	**40.0**	5.7
· · · Connotation	74.0	**78.0**	4.0
· · · Cause-effect	78.0	**88.0**	10.0
· · · Implication	68.5	**70.8**	2.3
· · · Part-whole	58.2	**70.9**	12.7
· · · Precondition	60.0	**65.7**	5.7
· · · Scenario	64.8	**76.5**	11.7
· · · Domain-specific*	13.3	**20.0**	6.7

Table 6: Average accuracy (%) on the annotated development set of C^3 per category ($\star$: only three instances).

6.6 The Usefulness of Contextualized Knowledge for Other Tasks

Instead of converting the extracted triples into weakly-labeled relation extraction instances and training from scratch, we simply 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 DialogRE. We compare the performance of methods that use different weights for parameter initialization except for the randomly initialized classification layer. We achieve $+2.9\%$ in F1 and $+3.1\%$ in $F1_c$ on DialogRE (Table 7). 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 relation extraction baseline. It might be interesting to investigate the relevance between document-level relation extraction and MRC for further performance boost.

parameter initialization	dev F1|$F1_c$	test F1|$F1_c$
(Yu et al., 2020)	65.5 (0.9) | 61.0 (0.4)	63.5 (0.9) | 58.7 (0.9)
–	64.9 (0.8) | 60.3 (0.7)	64.4 (1.1) | 59.2 (0.8)
{T0} in Table 3	66.4 (1.2) | 61.6 (0.8)	65.1 (1.9) | 60.3 (1.7)
{10} in Table 3	**67.1** (0.7) | **62.9** (0.8)	**67.3** (0.8) | **62.3** (0.7)

Table 7: Average F1 (%) and $F1_c$ (%) on DialogRE (–: RoBERTa-wwm-ext-large).

We also conduct preliminary experiments by fine-tuning the MRC student on multiple-choice translated MRC datasets DREAM (Sun et al., 2019b) and Cosmos QA (Huang et al., 2019) (originally in English) and see at least $+2.7\%$ in accuracy. As this is not the main focus of the paper, we provide more details in Appendix A.2.

7 Related Work

7.1 Contextualized Knowledge Extraction

A common solution to obtain external contextualized knowledge is to utilize existing knowledge bases via distant supervision (Ye et al., 2019). We extract contextualized knowledge from scripts, wherein contexts (i.e., scenes) are naturally aligned with verbal-nonverbal pairs to avoid noise. Besides, we focus on verbal-nonverbal knowledge to improve MRC, which is seldom studied.

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/semantic patterns (Zhang et al., 2020; Zhou et al., 2020) specific to each relation,

or lexical databases (Tandon et al., 2014, 2015). In contrast, we skip the step of naming relations between phrases and situate structured knowledge in its context. We extract knowledge based on language-independent, scenario-independent lexical patterns, which do not require dependencies such as training data, a broad-coverage semantic lexicon, or a high-quality syntactic parser.

7.2 Knowledge Distillation

Our teacher-student paradigm 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 data, each of our teachers is trained with **weakly-labeled data**. Due to the introduction of large-scale weakly-labeled data, we use the training data (hard/soft labels) of the target MRC task in all stages to guide the training. The model architecture of students is the **same** as that of teachers to improve the task performance, which is different from previous knowledge distillation studies (Li et al., 2014; Hinton et al., 2015).

8 Conclusions

This paper introduces how to extract contextualized verbal-nonverbal knowledge from scripts and use the knowledge to improve multiple-choice MRC. We compare different strategies to use the knowledge-based weakly-labeled data and further propose a multi-teacher paradigm to better learn from large-scale weakly-labeled data. Experimental results show the effectiveness of multi-teacher and usefulness of the knowledge for different tasks.

Acknowledgments

We would like to thank the anonymous meta-reviewers/reviewers for their insightful feedback. We would like to thank Prof. Ralph Grishman from New York University for his valuable suggestions.

References

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.

Yu-An Chung, Hung-Yi Lee, and James Glass. 2018. Supervised and unsupervised transfer learning for question answering. In *Proceedings of the NAACL-HLT*, pages 1585–1594.

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.

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.

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

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the NAACL-HLT*, pages 1066–1076.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the ACL*, pages 8342–8360.

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. 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.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*, cs.CL/1907.11692v1.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the EMNLP*, pages 2383–2392, Austin, TX.

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.

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.

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. 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. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint*, cs.CL/2107.02137v1.

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.

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*, pages 9233–9241.

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, 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.

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.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the ACL-IJCNLP 2021*, pages 4913–4922, Online.

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.

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

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. In *Proceedings of the Neurips*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the IJCAI*, pages 4004–4010.

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.

A Appendices

A.1 Result Analysis (II)

Besides the quantitative analysis provided in Section 6.5, we take a closer look at improved machine reading comprehension instances after we use the multi-teacher paradigm to leverage verbal-nonverbal knowledge ({10} vs. {T0} in Table 3). As shown in Example 1 in Table 8, the correct answer option *"eating something"* is not mentioned in the give dialogue. To answer this question correctly, we require external scenario knowledge about *"afternoon tea"*, which usually includes a mini-meal composed of sandwiches, cakes, etc. Injecting contextualized knowledge can power MRC models, perhaps because that scenes that compose a script are designed to clearly introduce the relevant location and temporal information (usually in scene headings) that may not be explicitly mentioned in the main body of the scenes.

In addition, nonverbal messages in action lines and parentheses provide additional information for utterances, which may help machine readers better "read between the lines". As shown in Example 2 in Table 8, by saying *"Don't think about it in this life."*, the male speaker implicitly rejected the suggestion of the woman on his bad temper.

Example 1	
F:	Hello, who would you like to speak to?
M:	Meimei, remember you asked me to have tea this afternoon? When will we meet?
F:	I'm sorry, I should have called you. You can decide it. I will be available this afternoon.
M:	Then at 3:30 in the afternoon, how about having afternoon tea at the place we last visited?
F:	Sure. I will be there on time.
Q	What will the woman and the man do in the afternoon?
A.	shopping.
B.	sing.
C.	eating something. ⋆
D.	seeing a movie.
Example 2	
F:	You really should control your bad temper!
M:	Don't think about it in this life.
Q	What can we know from this conversation?
A.	The woman can stand the temper of the man.
B.	The man is not ready to control his temper. ⋆
C.	The man let the woman control her temper first.
D.	The man accepts the woman's suggestion.

Table 8: English translation of improved machine reading comprehension instances from C^3 (⋆: the correct answer option; M: male; F: female; Q: question).

It might be an interesting future direction to explore how to map coarse-grained verbal-nonverbal knowledge into fine-grained types of general world knowledge such as CAUSE-EFFECT. However, it requires additional type-specific annotations.

A.2 The Usefulness of Contextualized Knowledge for Other Tasks (II)

Furthermore, as additional indications to evaluate the usefulness of the extracted knowledge, we use Google Translate to generate Chinese versions of DREAM and Cosmos QA, two popular multiple-choice MRC datasets for English in which most questions require general world knowledge. For the two translated multiple-choice MRC datasets, we simply use {10} in Table 3 to initialize an MRC model. Again, we report the average accuracy over five runs with different random seeds for all results.

data	type of construction	# of instances
C^3	human-annotated	19,577
DREAM	human-annotated	10,197
Cosmos QA	human-annotated	35,600

Table 9: MRC data statistics.

As shown in Table 10, in this noisy setting, we still obtain $+7.8\%$ in accuracy on the test set of DREAM and $+2.7\%$ in accuracy on the publicly available development set of Cosmos QA, by adapting our best-performing MRC student. The different performance levels on translated datasets and their original English versions may be due to the different sizes of text corpora for pre-training language models for English and Chinese and noise introduced by imperfect automatic machine translation. We will also release the code to extract contextualized knowledge from English scripts (Gorinski and Lapata, 2015) and leave exploiting English contextualized knowledge for MRC tasks or translating contextualized knowledge in other languages (such as Chinese) into English for future work.

parameter initialization	DREAM dev	DREAM test	Cosmos QA dev
RoBERTa-wwm-ext-large	61.4 (1.0)	60.8 (1.0)	56.7 (0.9)
{T0} in Table 3	67.0 (0.6)	65.5 (0.7)	57.1 (1.3)
{10} in Table 3	**69.2** (0.7)	**68.6** (1.0)	**59.4** (0.6)

Table 10: Average accuracy on the translated Chinese version of DREAM and Cosmos QA.

A.3 More Contextualized Verbal-Nonverbal Knowledge Triples

We provide knowledge triples extracted using each of the four types of patterns in Table 11.

type	nonverbal	verbal
Scene 3		
	The ball is lowered, lighting up a sign that reads "1972."	
Crowd(over television):	...1! Happy New Year!	
	The people in the bar cheer and kiss each other. They blow horns and toss confetti into the air.	
	Forrest looks around as Carla and Lenore lean over and kiss him.	
B_c	over television	Crowd: ...1! Happy New Year!
Scene 4		
Laifu Zhao:	Le Xiaocong, it is you! I thought there was a thief in the kitchen.	
Xiaocong Le:	Monitor, I got in without your permission. I'm really sorry.	
Laifu Zhao:	It's okay, it's okay. He handed a cup of hot water:	
	This room is too cold. Come here and take a sip of hot water to warm up. Take a break.	
Xiaocong Le took the cup of hot water:	Thank you! Monitor!	
Laifu Zhao:	Hey, Le Xiaocong, see how dedicated and focused you are. What are you writing?	
B_n	took the cup of hot water	Xiaocong Le: Thank you! Monitor!
Scene 5		
Wukong:	Huh! Why does this picture move?	
Aoguang:	This is a cruiser. The picture on the screen is a submarine fiber optic cable being laid.	
Wukong:	A cruiser?	
Aoguang:	Yes. With it, finding a needle in a haystack is not difficult. Hahaha...	
Wukong:	Finding Huaguoshan is easier than finding a needle in a haystack!	
Aoguang dumbfounded:	Well, well, well...	
B_n	dumbfounded	Aoguang: Well, well, well...
Scene 6		
	From a cauldron on the stove, hot water is poured into two pails, by the a kitchen boy under the nurse's command.	
Scullery Maid (O.S.):	Thomas Kent, sir? No sir.	
Will (O.S.):	The actor.	
Nurse:	Who asks for him?	
	Will has come to the kitchen door with a letter.	
Will:	William Shakespeare, actor, poet, and playwright of the Rose.	
	The nurse sends the scullery maid back to work.	
Nurse:	Master Kent is... my nephew.	
Will (giving her the letter):	I will wait.	
Nurse:	Much god may it do you.	
B_c	giving her the letter	Will: I will wait.
O	Will has come to the kitchen door with a letter.	Nurse: Who asks for him?
O	The nurse sends the scullery maid back to work.	Will: William Shakespeare, actor, poet, and playwright of the Rose.
Scene 7		
	Scene 90: outside the field	
Principal Qi:	(nodding his head) Excellent Game! Good job! Give students some water!	
Dean of Education:	Water is here! (distributed bottles of water one by one)	
Haiyin Lin:	Everyone played well in today's game!	
Gao Feng:	Unfortunately, we still lost by two points.	
	Several people nodded	
I	nodding his head	Principal Qi: Excellent Game! Good job! Give students some water!
I	distributed bottles of water one by one	Dean of Education: Water is here!
O	Several people nodded.	Gao Feng: Unfortunately, we still lost by two points.
Scene 8		
Mrs. Gump:	You do your very best now, Forrest.	
Forrest:	I sure will, Momma.	
Forrest (V.O.):	I remember the bus ride on the first day of school very well.	
Bus Driver:	Are you comin' along? Forrest: Momma said not to be taking rides from strangers.	
Bus Driver:	This is the bus to school.	
Forrest:	I'm Forrest Gump.	
Bus Driver:	I'm Dorothy Harris.	
Forrest:	Well, now we ain't strangers anymore.	
	The bus driver smiles as Forrest steps up into the bus.	
O	The bus driver smiles as Forrest steps up into the bus.	Forrest: Well, now we ain't strangers anymore.

Table 11: Examples of extracted verbal-nonverbal pairs situated in scenes (all translated into English).