Benchmarking Knowledge-Enhanced Commonsense Question Answering via Knowledge-to-Text Transformation

Ning Bian,[1,3] **Xianpei Han,**[1,2,*] **Bo Chen,**[1,2] **Le Sun**[1,2,*]

[1]Chinese Information Processing Laboratory, [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
{bianning2019, xianpei, chenbo, sunle}@iscas.ac.cn

Abstract

A fundamental ability of humans is to utilize commonsense knowledge in language understanding and question answering. In recent years, many knowledge-enhanced Commonsense Question Answering (CQA) approaches have been proposed. However, it remains unclear: (1) How far can we get by exploiting external knowledge for CQA? (2) How much potential of knowledge has been exploited in current CQA models? (3) Which are the most promising directions for future CQA? To answer these questions, we benchmark knowledge-enhanced CQA by conducting extensive experiments on multiple standard CQA datasets using a simple and effective knowledge-to-text transformation framework. Experiments show that: (1) Our knowledge-to-text framework is effective and achieves state-of-the-art performance on CommonsenseQA dataset, providing a simple and strong knowledge-enhanced baseline for CQA; (2) The potential of knowledge is still far from being fully exploited in CQA — there is a significant performance gap from current models to our models with golden knowledge; and (3) Context-sensitive knowledge selection, heterogeneous knowledge exploitation, and commonsense-rich language models are promising CQA directions.

Introduction

Using a variety of knowledge to help in understanding the meaning of language is one of the key abilities of humans (Minsky 2000). Commonsense question answering (CQA) evaluates whether machines can understand language like humans do by asking questions whose answers rely on commonsense knowledge. For example, Figure 1 shows a question, and the answer to this question needs commonsense knowledge "*puzzle is used for intellectual challenge*".

Witnessed the importance of commonsense knowledge for CQA, many studies have been conducted to incorporate external knowledge bases (KBs) in CQA models. These approaches usually leverage knowledge to enhance a specific CQA component: 1) enhancing representations (Weissenborn, Kočiský, and Dyer 2017; Bauer, Wang, and Bansal 2018; Mihaylov and Frank 2018; Ma et al. 2019); 2) enhancing attention mechanism (Chen et al. 2018; Wang and Jiang 2019); and 3) enhancing reasoning mechanism (Lin et al. 2019; Lv et al. 2020).
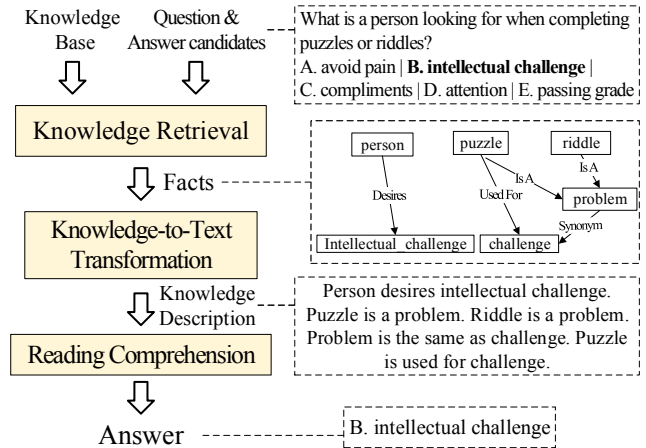
---

[*]Corresponding authors.

Figure 1: Our knowledge-to-text framework for benchmarking knowledge-enhanced CQA with an example from CommonsenseQA (Talmor et al. 2019).

Although many knowledge-enhanced CQA approaches have been proposed, we found it is still unclear: (1) How far can we get by exploiting external knowledge for CQA? (2) How much potential of knowledge has been exploited in current models? For example, can GNN-based models (Lin et al. 2019; Lv et al. 2020) encode and exploit all useful evidence provided by external knowledge? (3) Which are the most promising directions for knowledge-enhanced CQA? We believe answering these questions can provide valuable insights for future CQA studies and shed light on other knowledge-dependent tasks like reading comprehension (Rajpurkar et al. 2016) and conversation generation (Zhou et al. 2018).

To answer the above questions, we benchmark knowledge-enhanced CQA by conducting extensive experiments on multiple standard datasets via a simple and effective knowledge-to-text transformation framework. Intuitively, to benchmark knowledge-enhanced CQA, external knowledge should be incorporated in a simple way that is not specialized to specific models/components. This is challenging, due to 1) the heterogeneity between structured knowledge and unstructured textual questions/answers, i.e., knowledge facts are usually triples such as $<person, \textbf{Desires}, Intellectual\_challenge>$, but questions and answers are text; and 2) the context-

sensitivity of knowledge, i.e., a KB may contain thousands of facts about a concept, but only several of them are relevant to the given question. For example, among the thousands of facts about "*person*", only <*person, **Desires**, Intellectual_challenge*> is useful for answering the question in Figure 1.

Specifically, our knowledge-to-text framework consists of three stages, which are shown in Figure 1. Firstly, we retrieve facts from a commonsense knowledge graph (CKG). Then we transform the knowledge facts to textual descriptions via three transformation algorithms (template-based, paraphrasing-based, and retrieval-based). Finally, we utilize machine reading comprehension (MRC) models to predict answers by exploiting both the original questions and the textural knowledge descriptions. This framework is simple and general for benchmarking knowledge-enhanced CQA: 1) By transforming structured knowledge into textual descriptions, our method resolves the heterogeneity problem between knowledge and text. 2) By adopting MRC models, our method can learn to select question-relevant knowledge automatically. 3) Our simple knowledge-enhancing strategy allows us to easily compare the effects of different commonsense knowledge.

We conduct thorough experiments on multiple standard CQA datasets (Talmor et al. 2019; Levesque, Davis, and Morgenstern 2012; Zellers et al. 2019; Sap et al. 2019b).

The contributions of our paper are:

1. Through benchmarking experiments we found that the potential of external knowledge is still far from exploited in knowledge-enhanced CQA, i.e., current methods can only exploit knowledge to a limited extent. In our experiments, there is a big performance gap from current models to our models using golden knowledge.

2. We propose a simple and effective knowledge-to-text framework for knowledge-enhanced CQA which achieves state-of-the-art performance on the CommonsenseQA dataset, providing a simple and strong knowledge-enhanced baseline for CQA.

3. Our experimental results shed light on three important future directions for knowledge-enhanced CQA: context-sensitive knowledge selection, heterogeneous knowledge exploitation, and commonsense-rich language models.

Knowledge-enhanced CQA via Knowledge-to-Text Transformation

Following CommonsenseQA (Talmor et al. 2019), the CQA task in this paper is a multiple-choice problem with five answer candidates. Given question $Q = [q_1, q_2, \ldots, q_n]$ and answer candidates $A = \{A_1, A_2, \ldots, A_m\}$ with each answer candidate $A_k = [a_1^k, a_2^k, \ldots, a_l^k]$, a_j^k and q_i are words, i and j are indexes of words and k is the index of answer candidate, a CQA model needs to choose the correct answer from A.

We propose a simple and effective knowledge-to-text framework for benchmarking knowledge-enhanced CQA. Our framework includes three steps: 1) retrieving facts from CKG; 2) transforming knowledge to text; and 3) adopting an MRC model to select the answer.

Notice that the purpose of our paper is to benchmark knowledge-enhanced CQA rather than to propose new techniques. So, it is critical to select classical, robust, and well-known models, rather than new models which may lead to biased conclusions. Our framework is not specialized to a specific CQA setting, therefore it can also be used in other MRC or QA tasks.

In the following, we describe the three stages of our framework.

Knowledge Retrieval

To answer a question Q, our method first retrieves relevant knowledge from a given CKG. For example, to answer the question in Figure 1, we want to retrieve facts like <*person, **Desires**, Intellectual_challenge*> and <*puzzle, **UsedFor**, challenge*>. Following a previous study (Lin et al. 2019), we retrieve paths on CKG connecting question concepts and answer concepts as relevant facts, which provides a good precision/recall trade-off for question-relevant facts.

Concretely, given a question Q and an answer candidate A_k, we first identify concepts in them by exactly matching n-grams with the concepts in CKG. (we use ConceptNet (Speer, Chin, and Havasi 2017) in this paper). Then, for each pair of <*question concept, answer candidate concept*>, we find all paths between them on CKG (within K hops) as facts for A_k (K is a hyper-parameter here). For the example in Figure 1, "*puzzle→**IsA**→problem→**Synonym**→challenge*" is a 2-hop knowledge path for answer candidate "*intellectual challenge*".

Knowledge-to-Text Transformation

This section describes how to resolve the heterogeneity problem between knowledge and text via knowledge-to-text transformation. Specifically, we propose three transformation algorithms: template-based, paraphrasing-based, and retrieval-based, which are described as follows.

Template-based transformation. This algorithm transforms knowledge to text using a description template for each relation in a CKG. For example, we can use a template "X is a Y" to generate the description of <*puzzle, **IsA**, problem*> as "*puzzle is a problem*". Because the number of relations in a CKG is limited, we manually design a template for each relation type. For a knowledge path $\{k_1, k_2, \ldots, k_p, \ldots\}$ where k_p is a knowledge triple and p is its index, we sequentially generate a sentence for each tuple, i.e., $\{s_1, s_2, \ldots, s_p, \ldots\}$ where sentence s_p describes triple k_p.

Paraphrasing-based transformation. The main drawback of the template-based algorithm is the diversity issue, i.e., it always generates the same description for one relation. To address this issue, we employ a paraphrasing model to generate more diverse and fluent knowledge descriptions. Specifically, given the template-based description of a knowledge path, we generate its top-M paraphrases using beam-search decoding and concatenate them as the knowledge description. We adopt the encoder-decoder paraphrasing model trained on PPDB (Pavlick et al. 2015) and WikiAnswers (Fader, Zettlemoyer, and Etzioni 2013).

Retrieval-based transformation. The above two algorithms can only generate pseudo textual descriptions, which

Knowledge path	Template-based	Paraphrasing-based	Retrieval-based
Silk→**AtLocation**→China	Silk is located in China.	Silk is in China.	China is the world's largest silk producer.
Puzzle→**IsA**→Problem→**Synonym** →Challenge	Puzzle is a problem. Problem is the same as challenge.	Puzzles are problems. The problem is the same as the challenge.	Puzzle problem is a challenge game for children.
Walk←**MotivatedByGoal**←Hike→**Have-Subevent**→See beautiful views	Hike in order to walk. Hike have subevent see beautiful views.	You go hiking in order to go for a walk. You can see the beautiful scenery on hiking.	Burghclere has some beautiful rural scenery, so you can walk along the railway or go for a hike.

Table 1: Examples of knowledge descriptions generated by different knowledge-to-text algorithms.

are different from real-world knowledge descriptions. Therefore, we propose a retrieval-based knowledge-to-text algorithm, which retrieves texts from a real-world corpus (we use Wikipedia in this paper) as knowledge descriptions. Specifically, we adopt the distant supervision assumption (Mintz et al. 2009) that "*if a sentence contains the entities on a knowledge path, it will express the meaning of the knowledge path*". We split all Wikipedia documents into separate sentences and build a Wikipedia sentence retrieval system using Elastic Search. We use the knowledge descriptions from template-based transformation as queries to retrieve corresponding Wikipedia sentences containing the concepts on knowledge paths via the BM25 algorithm (Robertson and Walker 1994). Finally, the rank 1 sentence is used as the description.

To compare different knowledge-to-text transformation algorithms, Table 1 shows some examples of generated knowledge descriptions. We can see that: (1) The template-based algorithm can produce reasonable textual descriptions, although they may contain grammar errors (like "*Hike in order to walk*" in the 3rd example). (2) The paraphrasing-based algorithm can produce diverse and more fluent sentences ("*You go hiking in order to go for a walk*"), but may change some important words (e.g., "*beautiful view*" is changed to "*beautiful scenery*" in the 3rd example). (3) The retrieval-based algorithm can produce real-world sentences ("*China is the world's largest silk producer*") but may contain extra irrelevant content (like "*Burghclere*" in the 3rd example).

MRC-based Answer Prediction

Given a question and the generated knowledge descriptions, we predict its answer using MRC models. We adopt MRC models because: 1) MRC models can automatically learn to identify relevant information in a document (Seo et al. 2016). In our settings this ability can be used to automatically select question-relevant knowledge, as all knowledge facts have been transformed into a textual document; 2) MRC is a well-studied technique. Therefore, our method can directly leverage the strong ability of existing state-of-the-art MRC models, so that our benchmarking is effective, robust, and easy-to-implement.

Specifically, we model CQA as an MRC problem by treating knowledge descriptions as a document. In this way, current MRC models can be directly used, including BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), XL-Net (Yang et al. 2019b), and ALBERT (Lan et al. 2019) based MRC models. Figure 2 shows our MRC frame-
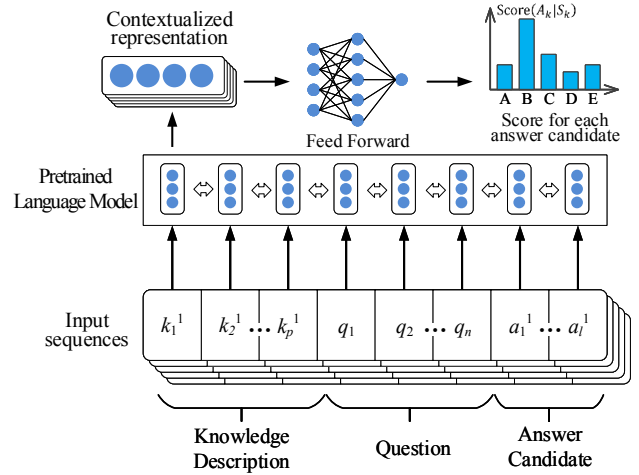

Figure 2: The MRC model for predicting answers in our knowledge-enhanced CQA method.

work. For each question, we construct a sequence $S_k = \{K_k, [SEP], Q, [SEP], A_k\}$ for each answer candidate A_k, where K_k is the generated knowledge descriptions, Q is the question, and $[SEP]$ is the separation token in pretrained language models (PLMs). Following Devlin et al. (2019), we use a feed-forward classifier as the output layer which predicts the answer score $\text{Score}(A_k|S_k)$. Finally, the highest-scored answer candidate is chosen as the answer.

Benchmarking Knowledge-Enhanced Commonsense Question Answering

This section benchmarks knowledge-enhanced CQA by conducting thorough experiments. We first verify the effectiveness and robustness of our knowledge-to-text-based CQA method, then we answer the three important questions: (1) *How far can we get by exploiting external knowledge for CQA?* (2) *How much potential of knowledge has been exploited in current models?* (3) *Which are the most promising directions for future knowledge-enhanced CQA?*

Experimental Settings

Datasets. We use CommonsenseQA dataset v1.11 (Talmor et al. 2019) as the primary dataset, and adopt the Winograd Schema Challenge (WSC, Levesque, Davis, and Morgenstern

Model	Knowledge Source	BERT	XLNet	RoBERTa	ALBERT
Human	–	88.9	88.9	88.9	88.9
Golden Knowledge	Human Explanations	81.1	85.1	84.7	83.7
Knowledge-to-Text					
Template-based	ConceptNet	67.9	77.5	78.1	81.1
Paraphrasing-based	ConceptNet	67.2	74.9	77.8	79.3
Retrieval-based	ConceptNet	65.0	75.0	77.1	79.4
Full	ConceptNet	**70.4**	**80.3**	**80.8**	**83.3**
Best Knowledge-enhanced System with Different PLMs	ConceptNet	69.0 (Ma et al. 2019)	79.3 (Lv et al. 2020)	80.8 (KEDGN)	(No available model so far)
Base Model	No knowledge	63.6	68.9	76.2	78.6

Table 2: Accuracies on CommonsenseQA.

2012), HellaSWAG (Zellers et al. 2019), and SOCIAL IQa (Sap et al. 2019b) as secondary datasets.

(1) *CommonsenseQA* (Talmor et al. 2019) contains 12,102 human-generated questions with 5 answer candidates for each question. All questions are elaborately designed to make sure commonsense knowledge is needed for correctly answering them. Furthermore, CoS-E (Rajani et al. 2019) provides each question with a human-annotated golden knowledge explanation. Due to the above advantages, We use CommonsenseQA as the primary benchmarking dataset.

(2) *WSC* (Levesque, Davis, and Morgenstern 2012) is a pronoun resolution dataset that requires commonsense knowledge, which is recognized as one of the most difficult CQA datasets (Zhou et al. 2020). Because WSC does not contain training data, we use WSCR (Rahman and Ng 2012) for training.

(3) *HellaSWAG* (Zellers et al. 2019) is an update of the commonsense reasoning dataset SWAG: given an event description like "*A woman sits at a piano*", a machine needs to select the most likely follow-up: "*She sets her fingers on the keys*". The "Overall accuracy" on the dev set is used in our evaluation.

(4) *SOCIAL IQa* (Sap et al. 2019b) is a QA dataset for commonsense reasoning about social situations, which requires emotional and social commonsense in a variety of every-day situations.

Knowledge base. We use ConceptNet 5 (Speer, Chin, and Havasi 2017) as the KB for benchmarking, because: (i) ConceptNet is general and can provide a large commonsense coverage for our CQA experiments. Other CKGs like ATOMIC (Sap et al. 2019a, if-then relations of events) and ASER (Zhang et al. 2020, relations of events, states, and actions) only contain partial knowledge for our experiments. (ii) The primary CommonsenseQA dataset is constructed upon ConceptNet and other datasets don't accompany a given KB. ConceptNet concepts can be easily and directly identified in questions and answers for CommonsenseQA, so that we can better benchmark knowledge-enhanced CQA by focusing on the ability of knowledge exploitation. We use the same 22 relations in ConceptNet as Talmor et al. (2019).

Baselines. We benchmark knowledge-enhanced CQA by assessing the performances of different MRC models

with/without external knowledge, including BERT-based (Devlin et al. 2019), RoBERTa-based (Liu et al. 2019), XLNet-based (Yang et al. 2019b), and ALBERT-based (Lan et al. 2019) MRC models.

To verify the effectiveness of knowledge-to-text transformation, we also report the performances of current knowledge-enhanced systems with corresponding pretrained language models as base encoders:

(1) Ma et al. (2019) (BERT + OCN + ConceptNet) is the best BERT-based knowledge-enhanced CQA system on CommonsenseQA, which uses an attention mechanism for knowledge incorporation and an Option Comparison Network (OCN) model for answer prediction.

(2) Lv et al. (2020) (XLNet + Graph Reasoning) is the best XLNet-based system on CommonsenseQA, which uses GNN to exploit knowledge from both ConceptNet and Wikipedia.

(3) KEDGN (RoBERTa + Knowledge) is the unpublished best RoBERTa-based knowledge-enhanced system on the leaderboard of CommonsenseQA, which exploits knowledge via a dual graph network. For a fair comparison, in Table 2 we report the accuracy of the best single model as described in its report.

Hyperparameters. For knowledge retrieval, we use knowledge paths within 2 hops ($K = 2$). In paraphrasing-based transformation, we use the top 1 paraphrasing result ($M = 1$). For MRC models, we initialize them with the official pretrained language models (BERT-Large, RoBERTa-Large, XLNet-Large, and ALBERT-XXLarge) and fine-tune them using CQA training data. The output layers have a 1024-dimensional hidden layer with a $tanh$ activation function. All models are trained using Adam with a learning rate of 5e-6.

Effect of Knowledge-to-Text Transformation

Table 2 and Table 3 show the experimental results on CommonsenseQA and other datasets. For our method, we use four settings: template-based, paraphrasing-based, retrieval-based, and a full model that uses a concatenation of all the three generated descriptions as a document. We found that:

1) *Knowledge-to-text transformation is effective for knowledge-enhanced CQA.* Our full model achieves state-of-the-art performance on CommonsenseQA. And all

Models	WSC	HellaSWAG	SOCIAL IQa
BERT	66.0	42.3	66.2
+ Knowledge	**68.1**	**44.2**	**68.8**
RoBERTa	81.4	82.5	74.3
+ Knowledge	**82.5**	**83.0**	**75.0**
ALBERT	84.9	86.1	77.2
+ Knowledge	**87.0**	**86.9**	**77.8**
Human	92.1	94.5	86.9

Table 3: Accuracies on other CQA datasets. "+ Knowledge" means using our knowledge-to-text transformation method (template-based) with 2-hop knowledge paths on ConceptNet. Human accuracy of WSC is reported by Bender (2015).

Path Length	BERT	XLNet	RoBERTa	ALBERT
1-hop	67.1	74.7	77.9	80.0
2-hop	**67.9**	**77.5**	**78.1**	**81.1**
3-hop	65.0	68.6	77.2	79.2

Table 4: Accuracies on different lengths of knowledge paths (template-based method).

template-based, paraphrasing-based, and retrieval-based models achieve improvements over non-knowledge base models.

2) *Knowledge-to-text transformation can robustly exploit knowledge for CQA.* Table 3 shows that our method can consistently improve the performances on three extra CQA datasets by exploiting external commonsense knowledge. Notice that although ConceptNet is not specially designed for WSC, HellaSWAG, and SOCIAL IQA datasets, our method can still achieve improvements, which further verifies the robustness of our method, and we believe the results on these datasets can be further improved if more relevant commonsense knowledge sources are available. In Table 2 our method achieves accuracy improvements on all base models (BERT, RoBERTa, XLNet, and ALBERT) and all settings (template-based, paraphrasing-based, and retrieval-based). Table 4 shows that our method is also robust on different lengths of knowledge paths, and the 2-hop knowledge path setting achieves the best performance.

3) *The three knowledge-to-text transformation algorithms are complements of each other.* In Table 2, the full model can achieve the best performance by combining all three knowledge-to-text algorithms, which verifies that these algorithms can complement each other. Among the three single algorithms, the template-based algorithm obtains the best performance. This may be because it is easier for MRC models to capture regularities in simple and formal sentences.

Overall, the above results verify that our simple knowledge-to-text transformation is a good strategy for benchmarking the effectiveness and robustness of knowledge-enhanced CQA.

In the following, we conduct benchmarking experiments on the primary CommonsenseQA dataset using the full model and 2-hop knowledge path setting.

Effect of Knowledge for CQA

This section studies "*how far can we get by exploiting external knowledge for CQA?*". To answer this question, Table 2 further shows the performances of MRC models using manually-annotated golden knowledge for each question (Rajani et al. 2019) as the knowledge description. We can see that:

By incorporating golden external knowledge, CQA can be significantly improved and can achieve close-to-human performance. On all BERT, XLNet, RoBERTa, and ALBERT-based MRC models, incorporating golden knowledge can significantly achieve 27%, 14%, 11%, and 7% accuracy improvements, correspondingly. The best golden-knowledge enhanced system (XLNet + Golden) can achieve 85.1% accuracy, which is not far from the human accuracy of 88.9%.

These results show that knowledge can get us quite far, and it is promising to study more effective knowledge-enhanced CQA models.

Effect of Knowledge in Current Models

This section investigates "*how much potential of knowledge has been exploited in current models?*". From Table 2, we can see that:

1) *Current knowledge-enhanced CQA methods only exploit knowledge to a limited extent.* In Table 2, we can see that: (i) compared with models using golden knowledge, all knowledge-enhanced CQA models have a big performance gap; and (ii) our simple knowledge-to-text strategy can achieve competitive performance with the complicated GNN-based strategies (KEDGN and XLNet + Graph Reasoning) and Option Comparison Network.

2) *Despite the effectiveness of our method, there is still great potential in generating accurate question-relevant knowledge descriptions.* To show this, Table 5 shows several bad cases of knowledge descriptions. We can see that, the golden knowledge descriptions are typically simple, relevant, and accurate, while the automatically generated descriptions may miss important evidence (1st example), be too complicated (2nd example), or contain noisy knowledge (3rd example). Based on these observations, we believe seeking and identifying more accurate question-relevant knowledge can further improve the knowledge exploitation ability of CQA methods.

3) *The commonsense knowledge embedded in current pretrained language models is still not enough for CQA.* In Table 2, we can see that there is a significant performance gap between base models without using knowledge and knowledge-enhanced models, although they have been trained using very large text corpus. To further study this, we also experiment using ERNIE (Zhang et al. 2019b), a knowledge-enhanced pretrained language model based on BERT, but the performance is lower than BERT-based models (60.0% accuracy on CommonsenseQA). We believe this is because ERNIE focuses on entity-centric facts, instead of commonsense. This shows that, although trained on very large text corpus, state-of-the-art pretrained language models still can not encode enough commonsense knowledge.

The above results show that the potential of knowledge

Missing Important Evidence	Question	*What could people do that involves **talking?***
	Answer candidates	***confession**	state park	sing	opera	carnival*
	Golden knowledge	***confession** involves **talking**.*
	Knowledge description	*people is located in confession. people is used for talk.*
Complicated Descriptions	Question	*They were getting ready for a really long **hike**, he put the food can in his what?*
	Answer candidates	***backpack**	make person sick	cabinet	house	recycling center*
	Golden knowledge	***backpacks** are used on **hicks**.*
	Knowledge description	*food can is located in backpack. **backpack is in the context of sport. hike is in the context of sport**......*
Noisy Knowledge	Question	*Most people who are **family** <u>like</u> to greet each other with a what?*
	Answer candidates	*listen to music	have friends	know what ophiolites	**hug**	apartments*
	Golden knowledge	*people who are family like to hug.*
	Knowledge description	*person desire hug. <u>person is located in **family**</u>. kissing have subevent **hug**. kissing cause <u>like</u>. meeting friend have subevent hug. hug in order to love. love is located in family. <u>most people desire hug.</u>*

Table 5: Bad examples of generated knowledge descriptions (template-based) and golden knowledge, where: 1) In the 1st example, the relational knowledge between "*talking*" and "*confession*" is missing in the generated knowledge description because it is not covered by ConceptNet. 2) In the 2nd example, knowledge description provides the knowledge about "*backpack*" and "*hike*" using two separate sentences, which is more complicated than golden knowledge and thus puts an extra burden on MRC models. 3) In the 3rd example, there are many irrelevant/noisy sentences in knowledge description about unimportant question words (like "*people*" and "*like*").
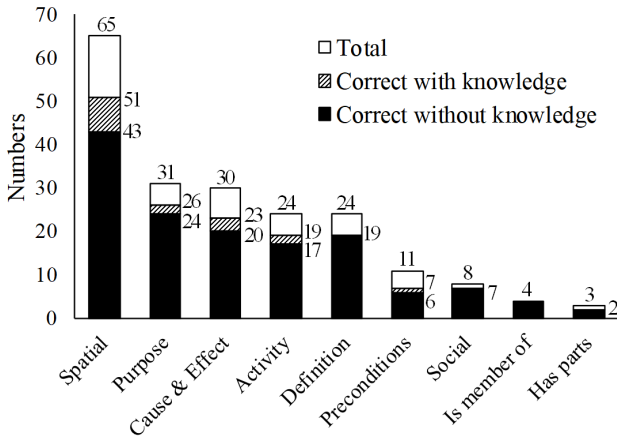

Figure 3: Performances of different commonsense skills using XLNet-based model, with/without knowledge descriptions (template-based).

is still far from being fully exploited by current knowledge-enhanced CQA methods. This is because of 1) the limited ability of current CQA models to exploit knowledge; 2) the lack of ability to identify accurate question-relevant knowledge; 3) the limited commonsense captured in pretrained language models.

Detailed Analysis

This section analyzes our method in detail.

Performances on Different Commonsense Skills. CQA questions require different types of commonsense skills (LoBue and Yates 2011). To analyze the effects of knowledge on different commonsense skills, we randomly sample 200 questions from CommonsenseQA and annotate their required

skills using the commonsense skill categories from Talmor et al. (2019).

Figure 3 shows the performances of our CQA method with/without knowledge on different skills. From Figure 3, we can see that: (1) Knowledge can significantly improve skills including "*Spatial*" (+12.3%), "*Cause & Effect*" (+10.0%), "*Activity*" (+8.3%) and "*Purpose*" (+6.5%). (2) For "*Definition*", "*Social*", and "*Has parts*" skills, the knowledge-enhanced model achieves similar performances with the base model. We believe this may be because ConceptNet has a low coverage for these types of knowledge.

Error Analysis. To understand why our model fails in some cases, we randomly select 50 error cases and group them into several categories. Table 6 shows the main error types with their examples:

1) *Indistinguishable knowledge*, i.e., retrieved knowledge cannot provide enough information for distinguishing answer candidates. For example, the 1st error case provides strong support for both correct and incorrect answers ("*airplanes can slow down/speed up*"). This is the main error type of our method (21 out of 50).

2) *Noisy knowledge*. Noisy knowledge misleads MRC models to give wrong answers, which often appears when knowledge descriptions are too long. In the 2nd error case, we can see that the important fact "*curtain is located in show*" is obscured by noisy facts about irrelevant concepts like "*seat*".

3) *No Knowledge*. Knowledge retrieval may not be able to retrieve question-relevant facts and thus provides no useful information for MRC models. From the 3rd case, we can see that the knowledge facts are all irrelevant to the answers.

The above three types of errors show that it is important to select accurate, complete, and context-sensitive knowledge for more effective knowledge-enhanced models.

Error type	Example	
Indistinguishable Knowledge (21/50)	Question	*What do **airplanes** do as they are arriving at the gate?*
	Answer candidates	$\sqrt{}$ ***slow down*** $\mid \times$ *land* $\mid \times$ *crash* $\mid \times$ ***speed up*** $\mid \times$ *carry people*
	Knowledge for correct answer	*airplanes can slow down.*
	Knowledge for predicted answer	*airplanes can speed up.*
Noisy Knowledge (15/50)	Question	*I took my <u>seat</u>, the **curtains** drew <u>back</u> and I <u>enjoyed</u> the what?*
	Answer candidates	$\times$ *auditorium* $\mid \times$ *theatre* $\mid \times$ ***movie*** $\mid \sqrt{}$ ***show*** $\mid \times$ *airplane*
	Knowledge for correct answer	***curtain is located in show**. cover is opposite to <u>back</u>. <u>seat</u> is located in show. show is located in opera. curtain is located in opera. person desire <u>enjoy</u>. curtain is located in theater. . . . . . .*
	Knowledge for predicted answer	*movie is located in theater. curtain is located in theater.*
No Knowledge (13/50)	Question	*Some animals can fly thanks to their lightweight hollow what?*
	Answer candidates	$\times$ *heads* $\mid \times$ ***tails*** $\mid \sqrt{}$ ***bodies*** $\mid \times$ *bones* $\mid \times$ *eyes*
	Knowledge for correct answer	*bones is located in person. person desire fly.*
	Knowledge for predicted answer	***[NO KNOWLEDGE FACT IS RETRIEVED]***

Table 6: Several error cases of XLNet-based model with template-based knowledge descriptions.

Related Work

Knowledge-enhanced CQA. Many studies have been proposed to exploit commonsense knowledge for CQA. Rajani et al. (2019) propose to train a GPT-based explanation generation model using manually labeled corpus, but it relies on extra human effort. KagNet (Lin et al. 2019) represents external knowledge as a graph and reasons via graph convolution and LSTM. Ma et al. (2019) incorporate knowledge with text-to-knowledge attention and adopt a BERT-based Option Comparison Network for answer prediction. Lv et al. (2020) propose a GNN-based reasoning model over A heterogeneous knowledge graph of both ConceptNet and Wikipedia sentences. Compared with these methods, our knowledge-to-text method exploits knowledge in a simple way and knowledge can be effectively used by the whole model.

Knowledge Exploitation in Neural Models. There are many studies which leverage external knowledge to enhance models on a variety of NLP tasks (Lin, Sun, and Han 2017; Yang and Mitchell 2017; An et al. 2018; Yang et al. 2019a; Logan et al. 2019; Chen, Sun, and Han 2018). Chen et al. (2018) leverage semantic relations in WordNet to enhance attention and inference abilities in the NLI task. Mihaylov and Frank (2018) apply key-value memory to represent commonsense facts and use word-to-knowledge attention for cloze-style MRC. Bauer, Wang, and Bansal (2018) propose a mutual information-based knowledge selection method and fuse knowledge using gated attention for multi-hop reasoning. Zhang et al. (2019a) propose an attention-based knowledge selection method for coreference resolution. ERNIE (Zhang et al. 2019b) and K-BERT (Liu et al. 2020) incorporate knowledge in pretrained language models, but mainly focus on entity-centric facts in KBs instead of commonsense.

Machine Reading Comprehension. In recent years, many effective end-to-end MRC models have been proposed, including BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), XLNet (Yang et al. 2019b) and ALBERT (Lan et al. 2019) based models. It has been proven that MRC models can effectively encode information in a document and find the most relevant information for answer prediction. In this paper, these abilities are utilized to select and exploit relevant knowledge for knowledge-enhanced CQA.

Conclusions and Future Work

We benchmark knowledge-enhanced CQA using a simple and effective knowledge-to-text transformation framework and provides a strong knowledge-enhanced baseline for CQA. By conducting thorough experiments, we found that: (1) Our knowledge-to-text framework is effective and robust for knowledge-enhanced CQA; (2) It is promising to incorporate knowledge in neural models for CQA; (3) The potential of knowledge is still far from being fully exploited — there is a large performance gap from current models to our models using golden knowledge.

The above results also shed light on the promising directions for knowledge-enhanced CQA:

1) **Context-sensitive knowledge selection is critical for knowledge-enhanced CQA.** According to the error analysis, more than 70% of errors are caused by noisy knowledge and indistinguishable knowledge.

2) **The knowledge-text heterogeneity is a critical bottleneck for exploiting the information from both knowledge and text.** We address this heterogeneity problem via simple knowledge-to-text transformation, and even such a simple strategy can outperform many knowledge-enhanced models like GNN-based and attention-based models. Therefore, we believe more advanced solutions for the heterogeneity problem will further improve CQA, e.g., uniform representation learning and joint graph representations.

3) **It is valuable to incorporate more commonsense in pretrained language models.** From our experiments, we can see that current state-of-the-art pretrained language models like BERT and XLNet still only encode limited commonsense knowledge. So, we believe commonsense-rich language models will provide valuable techniques and resources for CQA.

Acknowledgments

References

An, B.; Chen, B.; Han, X.; and Sun, L. 2018. Accurate Text-Enhanced Knowledge Graph Representation Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 745–755. New Orleans, Louisiana: Association for Computational Linguistics.

Bauer, L.; Wang, Y.; and Bansal, M. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4220–4230. Brussels, Belgium: Association for Computational Linguistics.

Bender, D. 2015. Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS*, 39–45.

Chen, B.; Sun, L.; and Han, X. 2018. Sequence-to-Action: End-to-End Semantic Graph Generation for Semantic Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 766–777. Melbourne, Australia: Association for Computational Linguistics.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Inkpen, D.; and Wei, S. 2018. Neural Natural Language Inference Models Enhanced with External Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2406–2417. Melbourne, Australia: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fader, A.; Zettlemoyer, L.; and Etzioni, O. 2013. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1608–1618. Sofia, Bulgaria: Association for Computational Linguistics.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* .

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2829–2839. Hong Kong, China: Association for Computational Linguistics.

Lin, H.; Sun, L.; and Han, X. 2017. Reasoning with Heterogeneous Knowledge for Commonsense Machine Comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2032–2043. Copenhagen, Denmark: Association for Computational Linguistics.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2020. K-BERT: Enabling Language Representation with Knowledge Graph. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence*, 2901–2908.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

LoBue, P.; and Yates, A. 2011. Types of Common-Sense Knowledge Needed for Recognizing Textual Entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 329–334. Portland, Oregon, USA: Association for Computational Linguistics.

Logan, R.; Liu, N. F.; Peters, M. E.; Gardner, M.; and Singh, S. 2019. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5962–5971. Florence, Italy: Association for Computational Linguistics.

Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; and Hu, S. 2020. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence*, 8449–8456.

Ma, K.; Francis, J.; Lu, Q.; Nyberg, E.; and Oltramari, A. 2019. Towards Generalizable Neuro-Symbolic Systems for Commonsense Question Answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 22–32.

Mihaylov, T.; and Frank, A. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 821–832. Melbourne, Australia: Association for Computational Linguistics.

Minsky, M. 2000. Commonsense-based interfaces. *Communications of the ACM* 43(8): 66–73.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference*

on Natural Language Processing of the AFNLP, 1003–1011. Suntec, Singapore: Association for Computational Linguistics.

Pavlick, E.; Rastogi, P.; Ganitkevitch, J.; Van Durme, B.; and Callison-Burch, C. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 425–430. Beijing, China: Association for Computational Linguistics.

Rahman, A.; and Ng, V. 2012. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 777–789. Jeju Island, Korea: Association for Computational Linguistics.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4932–4942. Florence, Italy: Association for Computational Linguistics.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Robertson, S. E.; and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, 232–241. Springer.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4463–4473. Hong Kong, China: Association for Computational Linguistics.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* .

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4444–4451.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158.

Wang, C.; and Jiang, H. 2019. Explicit Utilization of General Knowledge in Machine Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2263–2272. Florence, Italy: Association for Computational Linguistics.

Weissenborn, D.; Kočiský, T.; and Dyer, C. 2017. Dynamic integration of background knowledge in neural NLU systems. *arXiv preprint arXiv:1706.02596* .

Yang, A.; Wang, Q.; Liu, J.; Liu, K.; Lyu, Y.; Wu, H.; She, Q.; and Li, S. 2019a. Enhancing Pre-Trained Language Representations with Rich Knowledge for Machine Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2346–2357. Florence, Italy: Association for Computational Linguistics.

Yang, B.; and Mitchell, T. 2017. Leveraging Knowledge Bases in LSTMs for Improving Machine Reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1436–1446. Vancouver, Canada: Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800. Florence, Italy: Association for Computational Linguistics.

Zhang, H.; Liu, X.; Pan, H.; Song, Y.; and Leung, C. W.-K. 2020. ASER: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, 201–211.

Zhang, H.; Song, Y.; Song, Y.; and Yu, D. 2019a. Knowledge-aware Pronoun Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 867–876. Florence, Italy: Association for Computational Linguistics.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019b. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451. Florence, Italy: Association for Computational Linguistics.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*, 4623–4629.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020. Evaluating Commonsense in Pre-Trained Language Models. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence*, 9733–9740.