ArT: All-round Thinker for Unsupervised Commonsense Question Answering

Jiawei Wang[1,2], **Hai Zhao**[1,2*]
[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
wjw_sjt@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Without labeled question-answer pairs for necessary training, unsupervised commonsense question-answering (QA) appears to be extremely challenging due to its indispensable unique prerequisite on commonsense source like knowledge bases (KBs), which are usually highly resource consuming in construction. Recently pre-trained language models (PLMs) show effectiveness as an alternative for commonsense clues when they play a role of knowledge generator. However, existing work either relies on large-scale in-domain or out-of-domain labeled data, or fails to generate knowledge of high quality in a general way. Motivated by human thinking experience, we propose an approach of **A**ll-**r**ound **T**hinker (ArT) by fully taking association during knowledge generating. In detail, our model first focuses on key parts in the given context, and then generates highly related knowledge on such a basis in an association way like human thinking. Besides, for causal reasoning, a reverse thinking mechanism is especially added to further enhance bidirectional inferring between cause and effect. ArT is totally unsupervised and KBs-free. We evaluate it on three commonsense QA benchmarks: COPA, SocialIQA and SCT. On all scales of PLM backbones, ArT shows its brilliant performance and outperforms previous advanced unsupervised models. Our code is available at https://github.com/WangJW424/commonsenseQA-ArT.

1 Introduction

Commonsense question-answering (QA) has been a more challenging natural language understanding (NLU) task than conventional QA tasks, for it requires extra commonsense knowledge, which cannot be directly acquired from the given context, to make an appropriate answer (Niu et al., 2021).

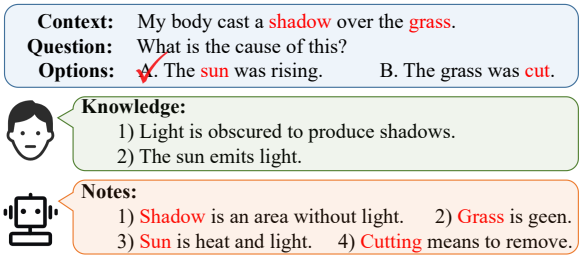

Figure 1: An commonsense QA exmaple in COPA. Text with blue background is the raw example and the gold answer is ticked. Text with green background is the needed commonsense knowledge. Text with orange background is notes generated by our ArT. Keyphrases are marked in red.

Figure 1 gives a typical example. To select the correct cause of the fact *"My body cast a shadow over the grass"*, models should know that *"Light is obscured to produce shadows"* and *"The sun emits light"*. As human beings, we can immediately make this association and judgment because above knowledge is deeply branded in our mind due to daily seeing and hearing, and maybe long-term education, which is so-called commonsense. However, it could be much difficult for models to be equipped with such ability. Besides, commonsense QA usually takes an unsupervised setting which makes this task even more difficult. This setting means there is no labeled training data available as commonsense is too broad to build a sufficient labeled training dataset (Shwartz et al., 2020).

To deal with it, prior work focused on building large-scale knowledge bases (KBs), also known as knowledge graphs (KGs), which usually contain millions of nodes and edges to record the relation between entities as relation triple: $<e_1, r, e_2>$ (Speer et al., 2017; Sap et al., 2019a). QA models then can be injected with commonsense through retrieving over KBs (Miller et al., 2016). Though impressive improvements are gained, such method is resource consuming in building or finding a

* Corresponding author. This paper was partially supported by Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Proceedings of the 29th International Conference on Computational Linguistics, pages 1490–1501
October 12–17, 2022.

good and suitable KB (e.g. Building ConceptNet needs 30 GB of RAM[1]). Recently, pre-trained language models (PLMs) have been widely used and proved to be effective in commonsense QA (Niu et al., 2021; Xia et al., 2022). Thanks to the self-supervised pre-training strategy on large-scale unlabeled text (e.g. WebText (Radford et al., 2019) and Wikipedia[2]), PLMs are competent for many tasks even under a zero-shot setting. And fine-tuning PLMs on task-specific data in a supervised way can further produce even stronger results (Schick and Schütze, 2021a; Gao et al., 2021; Schick and Schütze, 2021b). Since high-quality labeled datasets are rare, researching on unsupervised commonsense QA is still of great significance. With the help of PLMs, existing studies have explored some good solutions (Niu et al., 2021; Bosselut et al., 2021; Shwartz et al., 2020), however they either rely on large-scale in-domain or out-of-domain labeled data, or need to be specifically designed for different tasks. In this work, we focus on designing a simple and general method to solve commonsense QA tasks in a strictly unsupervised way.

Based on two empirical observations from human thinking,

(1) Given a question with context, people firstly tend to focus on several key parts (as marked in red in Figure 1) and then make corresponding associations to choose the right answer;

(2) For causal reasoning, people tends to carry out a bidirectional inferring to assist answer selection or verify the answer correctness.

we propose **A**ll-**r**ound **T**hinker (ArT) for unsupervised commonsense QA, which includes two principal methods: notes taking (NT) and reverse thinking (RT). Specifically,

(1) NT extracts some keyphrases out of the context and then generates corresponding notes, which will be added as extra knowledge in later evaluation. Based on an unsupervised keyphrase extractor, we designed our knowledge generation rule to be simple and general.

(2) RT converts the causal inferring question to two different forms: (*cause* $\rightarrow$ *effect*) and (*effect* $\rightarrow$ *cause*), and then integrates the decisions made from the two reverse directions.

[1]Declared by Speer et al. (2017) at https://github.com/commonsense/conceptnet5

[2]https://www.english-corpora.org/wiki

Our proposed model is strictly unsupervised and KBs-free for all it needs is PLMs. We test ArT and validate its effectiveness on three commonsense QA benchmark datasets: COPA (Roemmele et al., 2011), SocialIQA (Sap et al., 2019b) and SCT (Mostafazadeh et al., 2016). Our contribution is summarized as follows:

- ArT can generate highly related knowledge through the imitation of human behaviour and thought, which is qualified with inherent interpretablility.
- Compared with existing work, ArT is simple and general, which totally gets rid of the needs of any labeled data and the specific design on any specific task.
- We conduct experiments on 3 commonsense QA benchmarks with 4 different scales of PLMs, so as to reach solid and reproducible results. The results show that ArT outperforms other advanced unsupervised models.

2 Related Work

2.1 Building and Usage of Knowledge Bases

In order to equip QA models with commonsense reasoning ability, previous researches were devoted to building and retrieving large-scale knowledge bases (KBs). ConceptNet (Speer et al., 2017) is one of the most famous traditional KBs, which contains over 21 million edges (for relations) and over 8 million notes (for entities). While ATOMIC (Sap et al., 2019a) focuses more on *if-then* relations between events.

Previous work applies a relatively standard routine to solve commonsense QA. Given an existing KB, QA models can retrieve relation triples $<e_1, r, e_2>$ over it, which can be injected into models as word embedding directly (Wang et al., 2014; Paul and Frank, 2019) or first converted to a complete sentence according to preset templates (e.g. <bird, *CapableOf*, fly> $\rightarrow$ *Bird can fly.*) and then integrated with the raw input text (Ma et al., 2019; Mihaylov and Frank, 2018; Bauer et al., 2018). This type of methods may give remarkable performance for commonsense QA (Weissenborn et al., 2017). However, building and retrieving of KBs are both resource consuming (Bosselut et al., 2021).

Bosselut et al. (2019) claimed that commonsense knowledge does not cleanly fit into a schema of comparing two entities with a known relation so

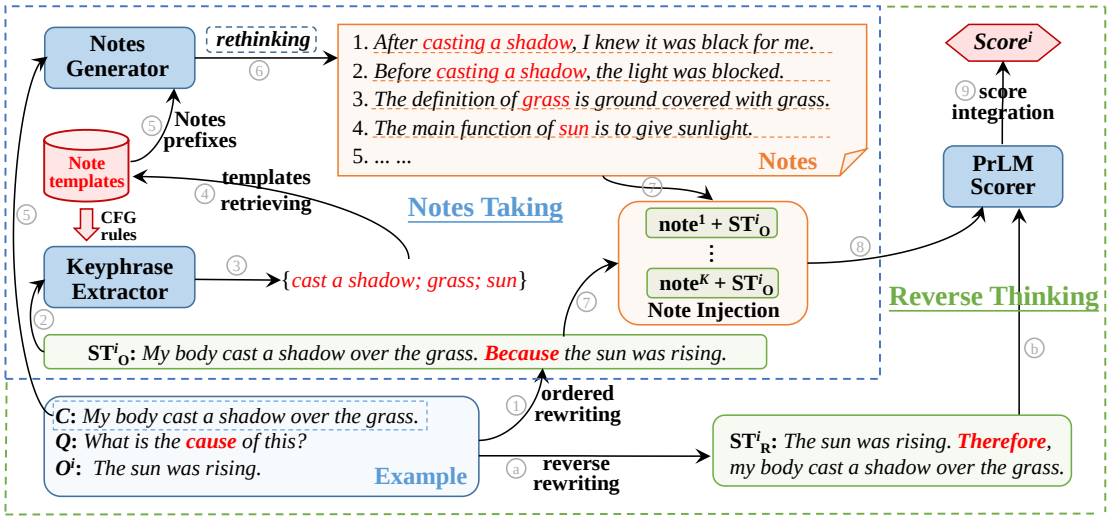

Figure 2: Overview of the proposed method **A**ll-**r**ound **T**hinker (ArT), which contains two principal method: Notes Taking (within blue dashline box) and Reverse Thinking (within green dashline box). The arrow denotes the flow of data stream, with number/letter within circle marking its order in time.

that they proposed COMET, which can generate rich and diverse commonsense descriptions in natural language. However, COMET has to extract knowledge triples from existing KBs as seed for PLM fine-tuning, while our method is totally KBs-free.

2.2 PLMs in Unsupervised Commonsense QA

Due to excellent performance and versatility, PLMs now dominate the backbone design of many NLP tasks, especially in QA (Wang et al., 2021; Zhang et al., 2021). Benefiting from the long-term pre-training on large-scale unlabeled text, PLMs are equipped with implicit or explicit commonsense (Davison et al., 2019). As a result, there comes a tendency to rely on PLMs as the sole source of world knowledge to solve commonsense QA in a zero-shot setting (Shwartz et al., 2020; Bosselut et al., 2021).

SEQA proposed by Niu et al. (2021) exploits generative PLMs to generate hundreds of pseudo-answers, and then applies Sentence-BERT (Reimers and Gurevych, 2019) to calculate semantic similarity between candidates and these generated pseudo-answers. A voting mechanism is designed to make final selection. Though SEQA shows impressive performance, it relies much on PLMs fine-tuned on large-scale labeled NLI (Natural Language Inference) datasets. Without such fine-tuning on such labeled datasets, SEQA's effectiveness will sharply decline. Instead, our method relies nothing but vanilla PLMs.

Shwartz et al. (2020) proposed Self-talk, which creatively applies PLMs as commonsense generator. It uses preset question prefixes to firstly prompt the PLM to generate information seeking questions (ISQs). ISQs will be put back to secondly prompt the PLM to generate their answers as clarifications, which will work as knowledge in later evaluation. Though this method is strictly unsupervised, the question prefixes are not general and have to be carefully designed according to different tasks. Besides, Self-talk fails to generate knowledge of high quality since the generation of ISQs and clarifications are both erratic. Inspired by Self-talk and towards its shortage, we designed a general method to generate knowledge of high relevance.

2.3 Unsupervised Keyphrase Extraction

Keyphrase extraction (KE) is the task of selecting several words or phrases that can summarize the main topic of the document (Hasan and Ng, 2014). Unsupervised KE methods use different features of the document, such as word frequency, position feature, relationship between words, etc (Mihalcea and Tarau, 2004; Bougouin et al., 2013).

SIFRank (Sun et al., 2020) is one of the most advanced unsupervised keyphrase extractors. It leverages different features of words in document and evaluates the weight of candidate phrases according to word embedding provided by PLMs. SIFRank is originally designed to extract key noun phrases from given documents, and we made slight modifications to adapt it to our model.

Key	Value	Replacing rule
"NP"	{ "*The definition of* [NP] *is*", "*The main function of* [NP] *is*", "[NP] *is a/an*" }	directly replace
"VP"	{ "[VP] *means*", "*After* [VP], ", "*Before* [VP], " }	convert to gerund first
"PNP"	{ "[PNP] *is a/an*", "[PNP] *felt*", "*After this,* [PNP]", "[PNP] *did this because*" }	directly replace

Table 1: Note templates lookup table.

3 All-round Thinker

Inspired by the behaviour and thought of human beings during solving questions requiring commonsense, we propose **A**ll-**r**ound **T**hinker (ArT). Figure 2 gives its overview. We will firstly introduce the task definition and the basic solution, and then describe the two principal methods of ArT in detail: Notes Taking (NT) and Reverse Thinking (RT).

3.1 Task Definition and Basic Solution

This work focuses on unsupervised commonsense QA task in multi-choice style, which consists of a *context* (C) with related *question* (Q), and asks models to select a single answer from a given *option* set: $O = \{O^i\}_{i=1}^{|O|}$. Though there are some variants of the task form, such as Piqa (Bisk et al., 2020) and WinoGrande (Sakaguchi et al., 2020), they can be conveniently transformed into this normalized formulation: (C, Q, O) (Shwartz et al., 2020).

Following previous work, we adopt a basic solution which uses PLMs as scorer. Based on the pre-training strategy: predicting the probability distribution of token n according to previous $(n-1)$-gram, e.g. OpenAI GPT (Radford et al., 2019), or bidirectional context, e.g. BERT (Devlin et al., 2019), PLMs are qualified for the role of *option* scorer even under a zero-shot setting (Bosselut et al., 2021). Sentence likelihood is the commonly-used scoring function for O^i:

$$\begin{aligned} \boldsymbol{S}(O^i|C,Q) &= P_{LM}(O^i|C,Q) \\ &= \frac{1}{|O^i|}\sum_{t=1}^{|O^i|}\log P_{LM}(O_t^i|C,Q,O_{<t}^i) \end{aligned} \tag{1}$$

where P_{LM} refers to the probability function abstracted from PLMs. The final predicted answer ($\hat{A}$) is selected as:

$$\hat{A} = \underset{O^i}{\operatorname{argmax}}\, \boldsymbol{S}(O^i|C,Q) \tag{2}$$

3.2 Notes Taking

To make full use of PLMs' potential of commonsense reasoning and overcome the shortage of Self-talk (Shwartz et al., 2020), we propose notes taking (NT) to generate commonsense descriptions in natural language (defined as "notes" in our work) in a simple and general way. NT is composed of two phases: keyphrase extraction and notes generation.

3.2.1 Keyphrase Extraction

Taking O^i and its corresponding C and Q as input (as shown in left bottom of Figure 2), ArT firstly rewrites the original interrogative Q into a declarative form (We followed the question rewriting method proposed by Shwartz et al. (2020)) and then concatenate $<C, Q, O^i>$ as a statement (ST_O^i). Then, we use an unsupervised keyphrase extractor to extract keyphrases from ST_O^i. To be specific, we extract three types of phrases: noun phrase (NP), verb phrase (VP) and person name phrase (PNP).

To implement this, for each type of phrase we designed a simple CFG (context-free grammar (Chomsky and Schützenberger, 1959)) rule to extract it out of the whole sentence, as following:

- $NP \rightarrow (nn|adj)*+nn$
- $VP \rightarrow vb + (pr)\{0,1\} + NP$
- $PNP \rightarrow (pn)\{1,2\}$

where *VP*, *NP* and *PNP* are non-terminators; vb (verb), nn (noun), adj (adjective), pn (person name) and pr (preposition) are terminators; '+' means concatenation; $\{a,b\}$ means repetition times range from a to b. '$*$' is equivalent to $\{0,\infty\}$.

Then, we add these CFG rules to the RegexpPaser tool of NLTK (Bird, 2006). We will extract top 5 most important keyphrases without any label or fine-tuned model but word embeddings obtained from a PLM, i.e. ELMo (Peters et al., 2018).

3.2.2 Notes Generation

Once keyphrases are obtained, we can retrieve a preset note templates set for getting notes prefixes. Considering that: (1) For an object (NP), people will think "what is it" and "what is it for"; (2) For a behavior (VP), people will think "what does it mean" and "what is the cause/effect"; (3) For a person (PNP), people will think "who is he/she"

and "what is his/her feeling/motivation/reaction", our proposed general-purpose note templates set is presented in Table 1.

We use the types of keyphrases as $keys$ and templates lists as $values$ to build the note templates set as a lookup table. Given such a lookup table, we can immediately retrieve the note templates for our extracted keyphrases and simply replace the tag ([NP], [VP] and [PNP]) with these keyphrases to form the note prefixes[3]. Though this lookup table seems to be simple, it shows effectiveness and generality on different benchmarks.

Next, for each note prefix, we will concatenate it to C and input them into a generative PLM to generate the complete note. Specifically, nucleus sampling (Holtzman et al., 2020) with $p = 0.8$ is applied as the decoding strategy rather than greedy/beam search to increase the **diversity** of generated text. Meanwhile, to ensure the **quality** and scale the number of generated notes, we sort all the notes according to their perplexity estimated by the PLM. We denote this process as ***Rethinking***. Finally, we will retain top K notes to construct the notes set: $\{note^k\}_{k=1}^{K}$, as shown in middle top of Figure 2. Each $note^k$ will be inserted into ST_O^i as extra knowledge to assist later *option* scoring. The score of O^i w.r.t $note^k$ is calculated as:

$$score_O^{i,k} = P_{LM}(O^i|note^k + ST_O^i - O^i)$$
$$= \frac{1}{|O^i|}\sum_{t=1}^{|O^i|}\log P_{LM}(O_t^i|note^k + ST_O^i - O^i + O_{<t}^i) \quad (3)$$

in which '+' means concatenation and '-' means removing. Eventually, a voting mechanism is applied to integrate the scores w.r.t all notes as:

$$Score_O^i = \frac{1}{K}\sum_{k=1}^{K} score_O^{i,k} \quad (4)$$

It is worth noting that there is no need in training from any labeled data throughout NT. And there is no need to modify the note templates according to different tasks since our note templates is designed for general purpose.

3.3 Reverse Thinking

For causal reasoning questions, we additionally introduce reverse thinking (RT) which conducts a bidirectional inferring between *cause* and *effect*. To implement this, besides the ordered rewriting (ST_O^i) as mentioned in Section 3.2.1, we also apply reverse rewriting that concatenates them in the order of $<O^i, Q_R, C>$ (denoted as ST_R^i), as shown in the right bottom of Figure 2. Note that Q_R is the opposite question of Q. To be specific, after question rewriting, "*Because*" and "*Therefore*" are two opposite questions in causal reasoning tasks.

To conduct bidirectional inferring, except for $Score_O^i$ as introduced in Section 3.2 for the ordered inferring: $C + Q \rightarrow O^i$, we set another scoring function $Score_R^i$ for reverse inferring: $O^i + Q_R \rightarrow C$, as:

$$Score_R^i = P_{LM}(C|ST_R^i - C)$$
$$= \frac{1}{|C|}\sum_{t=1}^{|C|}\log P_{LM}(C_t|ST_R^i - C + C_{<t}) \quad (5)$$

To take advantage of bidirectional inferring, we design a mixed scoring function by simply compute the average value of the above two:

$$Score_X^i = \frac{1}{2}(Score_O^i + Score_R^i) \quad (6)$$

Finally, formula (2) is applied to select the answer by replacing $\boldsymbol{S}$ with $Score_O^i$ (default) or $Score_X^i$ (for causal reasoning). From the perspective of model enhancing, averaging $Score_O^i$ and $Score_R^i$ can be regarded as the assembly of two models ("cause" model and "effect" model), which is a common method to enhance model performance and robustness.

4 Experiment

4.1 Datasets

ArT is evaluated on three different commonsense QA benchmarks: COPA(Roemmele et al., 2011), SocialIQA(Sap et al., 2019b) and SCT(Mostafazadeh et al., 2016). Here are the detailed information:

- **COPA**[4] (**C**hoice **o**f **P**lausible **A**lternatives): evaluates the ability of causal reasoning about a certain event, which is described as a single sentence. Each question is accompanied with two candidate options.

[3] In order to keep a correct grammar, for VP we will firstly convert its verb into gerund form before replacing.

[4] https://people.ict.usc.edu/~gordon/copa.html

- **SocialIQA**[5] (**Social** **I**nteraction **Q**uestion **A**nswering): evaluates the reasoning ability on social interactions. It has various questions, including the subject's motivation, reaction, personality, etc. Each question is accompanied with three candidate options.

- **SCT** [6] (**S**tory **C**loze **T**est): requires models to select the right ending of the given short story from two alternatives. Each story is composed of four sentences.

Since two test sets of the three datasets are hidden, we report all results on dev sets. Note that the labels are kept invisible and only used for final accuracy evaluating.

4.2 Baseline and Contrastive models

Our baseline is constructed as only using PLMs as scorer without any explicit knowledge injection. Formula (1) is applied as the scoring function as described in Section 3.1. We also compare ArT with other advanced unsupervised models:

- **Self-talk**(Shwartz et al., 2020): It acquires knowledge through a two-stage prompting of PLMs. Different question prefixes had to be specially designed for different tasks.

- **SEQA**(Niu et al., 2021): It applies PLMs to generate hundreds of pseudo-answers and compares them with each $option$. However, its scorer relies on PLMs fine-tuned on large-scale labeled NLI datasets, which is not strictly unsupervised. For fair comparison, we design another setting that replaces the fine-tuned PLM with the original one (only pre-trained on unlabeled text) .

- **CGA**(Bosselut et al., 2021): It employs a generative KB COMET(Bosselut et al., 2019), which is trained on an existing seed KB (e.g. ConceptNet), to construct context-relevant knowledge graphs to reason over.

4.3 Setup

Following previous work, we employ OpenAI GPT(Radford et al., 2019) as the PLM backbone. To reach solid and reproducible results, we conduct experiments on GPT-2 of 4 different scales: distil, medium, large and xlarge. For ArT and Self-talk, the same scale GPT is applied during both knowledge generating and $option$ scoring. For SEQA, GPTs of different scales are used for pseudo-answers generation and $\text{SRoBERTa}_{\text{large}}$(Reimers and Gurevych, 2019) is used for semantic similarity calculation. To distinguish, $\text{SRoBERTa}^{NLI}_{\text{large}}$ and $\text{SRoBERTa}^{Origin}_{\text{large}}$ refer to $\text{SRoBERTa}_{\text{large}}$ with and without further fine-tuning on NLI datasets, respectively. For Self-talk[7] and SEQA[8], we re-run their codes with their original settings and report both our re-running results[9] and results coming from their publications. For CGA, we report results provided by Niu et al. (2021). For ArT, we modified the open source code of SIFRank[10] to enable it to extract more kinds of phrases rather than only noun phrase. The size of notes set (K) is set as 32 as default. Except for $Score^i_O$, ArT takes another setting $Score^i_X$ on COPA.

4.4 Results

Table 2 shows the results on three benchmarks. The results of our re-running is highly consistent with those reported in their publications (last column). Note that published results on COPA seem to have a deviation with our reproduction. It because that they are on test set, while ours on dev set. Shwartz et al. (2020) reported 66.0% on COPA in their paper, which is tested on the dev set of a small version which contains 1/5 instances of that other researches used. And ArT reaches 68.0% on that set under the same setting.

On all datasets, ArT obtains state-of-the-art performance among almost all fully unsupervised models. Besides, it is noticed that on GPT of different scales, ArT with NT stably brings positive improvement over baseline. On causal reasoning task COPA, adding RT will bring further accuracy improvement.

In contrast with ArT, Self-talk fails to maintain effectiveness, whose accuracy is even slightly lower than baseline from time to time, especially on SCT. It indicates that the knowledge generated by Self-talk could be noisy and as a result it misguides model evaluation.

With the help of $\text{SRoBERTa}^{NLI}_{\text{large}}$, SEQA can reach very impressive results on all the datasets, especially SCT (exceed all models than over 10%).

[5] https://leaderboard.allenai.org/socialiqa

[6] https://www.cs.rochester.edu/nlp/rocstories/

[7] https://github.com/vered1986/self_talk

[8] https://github.com/heyLinsir/Semantic-based-QA

[9] For each setting except $\text{GPT-2}_{\text{xlarge}}$ (limited by the computational power), we run 3 times and report the average number.

[10] https://github.com/sunyilgdx/SIFRank

Dataset	Models	Our (re-)running				Published
		DistilGPT-2	GPT-2_{medium}	GPT-2_{large}	GPT-2_{xlarge}	GPT-2_{xlarge}
COPA	Baseline	57.8	62.4	65.8	66.0	–
	SEQA	51.4 (63.0)	53.0 (68.4)	53.8 (72.0)	54.4 (75.4)	79.4
	Self-talk	59.8 (↑2.0)	65.0 (↑2.6)	66.6 (↑0.8)	66.2 (↑0.2)	68.6
	CGA	–	–	–	–	72.2
	ArT	60.2 (↑2.4)	64.8 (↑2.4)	67.0 (↑1.2)	67.6 (↑1.6)	–
	ArT ($Score^i_X$)	**61.0** (↑3.2)	**65.6** (↑3.2)	**69.4** (↑3.6)	**69.8** (↑3.8)	–
SocialIQA	Baseline	41.3	44.3	45.5	45.9	–
	SEQA	34.9 (43.9)	35.9 (44.6)	36.5 (46.6)	36.6 (47.5)	47.5
	Self-talk	40.5 (↓0.8)	44.8 (↑0.5)	46.1 (↑0.6)	47.2 (↑1.3)	47.5
	CGA	–	–	–	–	45.4
	ArT	**42.0** (↑0.7)	**45.6** (↑1.3)	**47.6** (↑2.1)	**47.3** (↑1.4)	–
SCT	Baseline	59.6	67.4	69.1	70.5	–
	SEQA	50.7 (74.7)	53.3 (80.5)	54.2 (82.4)	54.9 (83.2)	83.2
	Self-talk	59.8 (↑0.2)	**68.5** (↑1.1)	69.2 (↑0.1)	70.4 (↓0.1)	70.4
	CGA	–	–	–	–	71.5
	ArT	**60.2** (↑0.6)	68.3 (↑0.9)	**69.5** (↑0.4)	**71.6** (↑1.1)	–

Table 2: Accuracy (%) on COPA, SocialIQA and SCT. All results except last column are run by ourselves. Best results are depicted in boldface (only consider fully unsupervised models for fairness). ↑/↓ refer to relative increase/decrease compared with baseline. For SEQA, we list results of two settings: SRoBERTa$^{Origin}_{\text{large}}$ (before brackets) and SRoBERTa$^{NLI}_{\text{large}}$ (in brackets).

However, when working in a strictly unsupervised mode, this method quickly becomes invalid (close to random selection).

We also observe a common phenomenon on all our ArT, baseline and contrastive models: when enlarging GPT-2 from large (750M parameters) to xlarge (1500M parameters), we encounter no obvious model performance increasing and even decreasing in some scenarios. It indicates that there could be a limit to only using the method of increasing model parameters to improve the performance of language models as sentence scorer or knowledge generator.

5 Analysis and Discussion

5.1 Effect of Different Modules

In order to determine the source of performance growth, we conduct ablation study on ArT, as shown in Table 3. As expected, injecting knowledge with our NT has positive effect on all three tasks. When removing the *Rethinking* mechanism, all results slightly decrease, which indicates that *Rethinking* can help increase the quality of generated knowledge. Besides, $Score^i_X$ can bring further improvement whether or not NT is employed on COPA.

Models	COPA	SocialIQA	SCT
Baseline	65.8	45.5	69.1
+$Score^i_X$	68.9	–	–
+NT	67.0	**47.6**	**69.5**
+NT−*Rethinking*	66.2	46.7	69.4
+NT+$Score^i_X$	**69.4**	–	–

Table 3: Ablation study for ArT modules on GPT-2_{large}.

5.2 Number of Notes and Keyphrases

Figure 3 shows the effect of notes set size K. On all the benchmarks, along with the increasing of K, the accuracy curves basically show an upward trend, which indicates that as the number of generated knowledge increases, ArT will not accumulate too much noise as Niu et al. (2021) observed in Self-talk. Therefore, compared with Self-talk, ArT tends to generate highly related knowledge, which is contributed by our NT mechanism.

In our experiments (Section 4), the number (N) of keyphrases to extract is set as 5 as default. To show the effect of N, we conduct an ablation study on SCT by setting $N \in \{1, 3, 5, 7, 9\}$. The PLM is selected as GPT-2_{distil}. Table 4 shows the results. It is noticed that extracting more keyphrases does

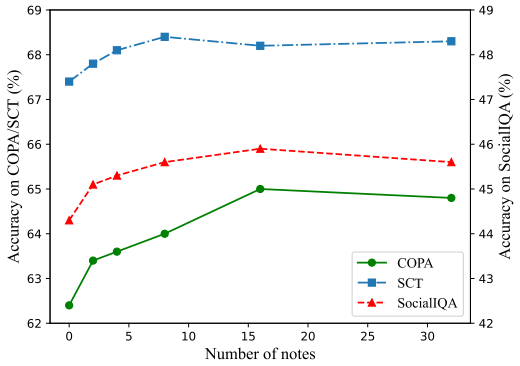

Figure 3: Accuracy curves of ArT on GPT-2_{medium} w.r.t the size K of notes set.

N	1	3	5	7	9
baseline			59.6		
ArT	59.6	**60.2**	**60.2**	60.0	60.0

Table 4: Accuracy (%) of ArT (GPT-2_{distil}) on SCT under different settings of N.

not always results in a better performance. But in general, the choice of N does not have an obvious impact on the final performance.

Dataset	Model	Rationality	Usefulness
COPA	Self-talk	0.24	0.20
	ArT	**0.32**	**0.28**
SocialIQA	Self-talk	0.17	0.16
	ArT	**0.26**	**0.27**

Table 5: Human evaluation on the rationality and usefulness of generated knowledge.

5.3 Quality of Generated Knowledge

To compare the quality of generated knowledge (whether the knowledge is reasonable enough to be a "fact" and correlative enough to be useful), we conduct human evaluation on the rationality (-1: *irrational*, 0: *meaningless*, 1: *rational*) and usefulness (-1: *negative*, 0: *neutral*, 1: *positive*) of knowledge generated by ArT and Self-talk. Two annotators are asked to independently annotate 100 randomly selected knowledge for both COPA and SocialIQA. The overall average scores of two annotators for each dataset are shown in Table 5. It is noticed that ArT outperforms Self-talk in generating knowledge with both rationality and usefulness.

We further make statistics on the kinds of all the knowledge generated by Self-talk and ArT on COPA. We divide them into four classifications:

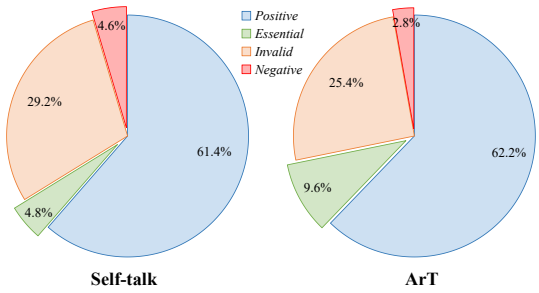

Figure 4: Statistics on the quality of knowledge generated by Self-talk and ArT on COPA.

- *Positive*: Baseline makes the right prediction. Adding the knowledge still makes a right prediction.
- *Essential*: Baseline makes the wrong prediction. Adding the knowledge helps make a right prediction.
- *Invalid*: Baseline makes the wrong prediction. Adding the knowledge still makes a wrong prediction.
- *Negative*: Baseline makes the right prediction. Adding the knowledge leads to a wrong prediction.

As shown in Figure 4, by doubling *Essential* and reducing *Negative* knowledge, ArT outperforms Self-talk in generating high-quality commonsense.

Figure 5 and 6 in Appendix show some examples of knowledge generated by ArT, Self-talk and SEQA for COPA, SocialIQA and SCT. As can be seen, ArT generates highly related knowledge on the basis of focusing on keyphrases of given context. While Self-talk may generate meaningless or even noisy knowledge. It is also noticed that SEQA could generate very reasonable pseudo-answers. And to distinguish the relationship between these pseudo-answers and given *options* relies much on a sentence embedding extractor fine-tuned on labeled NLI data.

5.4 Effect of Note Types

As we designed three types of notes: NP, VP and PNP, we only consider one type at one time to show the effect of each in different benchmarks, as shown in Table 6 (The PLM is GPT-2_{medium}).

It is noticed that each type of notes has a positive effect when applied separately, and their combination works better. Note that COPA has no person

Dataset	NP	VP	PNP	All / None
COPA	63.6	63.2	62.4	64.8 / 62.4
SocialIQA	45.0	44.8	45.3	45.6 / 44.3
SCT	68.0	67.9	68.1	68.3 / 67.4

Table 6: Effect of each type of notes on different tasks.

name phase (PNP), so PNP notes does not work on it. On SocialIQA, PNP works best among three types of notes. This is reasonable since SocialIQA focus much on human behavior in social interactions.

6 RT for Other Questions

Although RT is designed to enhance causal reasoning, we also explored if RT has the potential to help other questions. To investigate this, we apply it on other two datasets: SocialIQA and SCT. Note that in these tasks the opposite question is hard to define, therefore we simply exchange the position of $option$ and $context$, that is $ST_O^i =< C, Q, O^i >$ and $ST_R^i =< O^i, Q, C >$. We employ three scoring functions on basis of GPT-2_{medium} baseline, as shown in Table 7.

Functions	COPA	SocialIQA	SCT
$Score_O^i$	62.4	**44.3**	**67.4**
$Score_R^i$	63.2	42.5	62.8
$Score_X^i$	**65.3**	44.1	65.4

Table 7: Accuracy (%) of GPT-2_{medium} baseline with different scoring functions.

On all the tasks, $Score_O^i$ and $Score_R^i$ can obtain positive results (much better than random selection). By simply averaging ($Score_X^i$), on COPA it can reach a higher score. On other tasks, it falls into the middle of $Score_O^i$ and $Score_R^i$. Considering that $Score_R^i$ shows a comparable performance with $Score_O^i$ by simply exchanging the position of $option$ and $context$, developing a general method for opposite question definition or designing a more exquisite method to integrate $Score_O^i$ and $Score_R^i$ perhaps could make RT suitable for questions beyond causal reasoning, which will be the key point of our future work.

7 Conclusion

Commonsense QA has been a challenging task for it requires extra knowledge beyond the given context. In consideration of the high resource consumption of building knowledge bases (KBs) and the rarity of high-quality labeled data, this work aims at addressing commonsense QA in a fully KBs-free and unsupervised way. Inspired by the association process of human thinking, we propose **A**ll-**r**ound **T**hinker (ArT), which first focuses on key parts in the given context, and then generates highly related knowledge on such a basis in an association way. Besides, a reverse thinking mechanism is introduced to further enhance bidirectional inferring for causal reasoning as human will do. We test ArT on three benchmarks: COPA, SocialIQA and SCT. ArT outperforms previous advanced unsupervised models and shows stable performance on all scales of PLM backbones.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7432–7439.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.

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 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551.

Noam Chomsky and Marcel P Schützenberger. 1959. The algebraic theory of context-free languages. In *Studies in Logic and the Foundations of Mathematics*, volume 26, pages 118–161. Elsevier.

Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. 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)*, pages 1173–1178.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1262–1273.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. *EMNLP 2019*, page 22.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Todor Mihaylov and Anette Frank. 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 (Volume 1: Long Papers)*, pages 821–832.

Alexander H Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Yilin Niu, Fei Huang, Jiaming Liang, Wenkai Chen, Xiaoyan Zhu, and Minlie Huang. 2021. A semantic-based method for unsupervised commonsense question answering. In *ACL-IJCNLP 2021*.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*, pages 2227–2237.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8732–8740.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 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)*, pages 4463–4473.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4615–4629.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Yi Sun, Hangping Qiu, Yu Zheng, Zhongwei Wang, and Chaoran Zhang. 2020. Sifrank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, 8:10896–10906.

Jiawei Wang, Hai Zhao, Yinggong Zhao, and Libin Shen. 2021. What if sentence-hood is hard to define: A case study in Chinese reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2348–2359. Association for Computational Linguistics.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1591–1601.

Dirk Weissenborn, Tomáš Kočiskỳ, and Chris Dyer. 2017. Dynamic integration of background knowledge in neural nlu systems.

Mengzhou Xia, Mikel Artetxe, Jingfei Du, Danqi Chen, and Ves Stoyanov. 2022. Prompting electra: Few-shot learning with discriminative pre-trained models. *arXiv preprint arXiv:2205.15223*.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

		COPA	SocialIQA
Instance	**C:**	*The woman hired a lawyer.*	*Carson was excited to wake up to attend school.*
	Q:	*What is the cause of this?*	*Why did Carson do this?*
	O:	a) *She wanted to sue her employer.* b) *She decided to run for office.*	a) *Take the big test.* b) *Just say hello to friends.* c) *Go to bed early.*
ArT	**top-1:**	*After hiring a lawyer, she sued the man for sexual harassment.*	*Carson felt great and enjoyed being around people again.*
	top-2:	*Before hiring a lawyer, she has little chances to win in the court.*	*Attending school means working hard and learning something new everyday.*
Self-talk	**top-1:**	*Before [b)], the prosecutor said her client was a criminal.*	*Carson did this because they wanted him in the classroom.*
	top-2:	*"Lawyer" anyway is the word that I'm using to describe it.*	*Carson did this because they wanted him there for his education.*
SEQA	**top-1:**	*She thought the employer was trying to rip her off.*	*To focus on his studies.*
	top-2:	*She was worried her employer would be fired.*	*To play football with his friends and play street basketball.*

Figure 5: Top two most contributing generated knowledge for instances of COPA and ScocialIQA. The correct options are underlined. Keyphrases extracted by ArT are marked in red.

		SCT
Instance	**C:**	*Rick grew up in a troubled household. He never found good support in family, and turned to gangs. It wasn't long before Rick got shot in a robbery. The incident caused him to turn a new leaf.*
	Q:	*What is the ending of the story?*
	O:	a) *Rick is happy now.* b) *Rick joined a gang.*
ArT	**top-1:**	*After turning a new leaf, he decided that fighting crime was more important than robbing banks.*
	top-2:	*Turning a new leaf means not committing any crimes again.*
Self-talk	**top-1:**	*Rick is happy now means that he has learned from his mistakes.*
	top-2:	*Rick is happy now means that he can get his life back on track.*
SEQA	**top-1:**	*He was now a good person.*
	top-2:	*His life has dramatically improved with a period of rehabilitation and good mental health.*

Figure 6: Top two most contributing generated knowledge for instances of SCT. The correct options are underlined. Keyphrases extracted by ArT are marked in red.