Improving Unsupervised Commonsense Reasoning Using Knowledge-Enabled Natural Language Inference

Canming Huang, Weinan He, Yongmei Liu*
Dept. of Computer Science, Sun Yat-sen University, Guangzhou 510006, China
{huangcm, heweinan}@mail2.sysu.edu.cn, ymliu@mail.sysu.edu.cn

Abstract

Recent methods based on pre-trained language models have shown strong supervised performance on commonsense reasoning. However, they rely on expensive data annotation and time-consuming training. Thus, we focus on unsupervised commonsense reasoning. We show the effectiveness of using a common framework, Natural Language Inference (NLI), to solve diverse commonsense reasoning tasks. By leveraging transfer learning from large NLI datasets, and injecting crucial knowledge from commonsense sources such as ATOMIC 2020 and ConceptNet, our method achieved state-of-the-art unsupervised performance on two commonsense reasoning tasks: WinoWhy and CommonsenseQA. Further analysis demonstrated the benefits of multiple categories of knowledge, but problems about quantities and antonyms are still challenging.

A WinoWhy Example
WSC Question: Joan made sure to thank Susan for all the help she had received. She refers to Joan because **Reason**: Joan is doing the thanking so she must have received the help. **Label**: Positive
Convert WinoWhy to NLI
Premise: Joan is doing the thanking so she must have received the help. **Hypothesis**: Joan made sure to thank Susan for all the help Joan had received. **Label**: entailment

Table 1: A WinoWhy example consists of WSC question and reason, while the label is “Positive” or “Negative”. We use NLI as a common task and convert WinoWhy to NLI form.

1 Introduction

Recently, the task of commonsense reasoning has attracted much attention, as believed to be a critical and yet challenging component of human-level intelligence (Levesque et al., 2012, Davis, 2017; Wang et al., 2019a). To test models’ ability to understand natural language and reason with external commonsense knowledge, efforts have been made towards building many challenging WSC-like (Winograd Schema Challenge) tasks and QA (question-answer) tasks. Specifically, (Zhang et al., 2020a) crowd-sourced human-provided justifications as reasons for the WSC problems, resulting in a new dataset called WinoWhy. An example of WinoWhy is shown in Table 1, the model is asked to determine whether the given reason for the WSC problem is correct. Meanwhile, constructed based on ConceptNet (Speer et al., 2017), ComonsenseQA (Talmor et al., 2019) is designed as a five-choice QA dataset that requires model to capture the relation between the question and the correct answer. In this work, we experiment on these two commonsense datasets.

Although diverse methods based on pre-trained language models and external knowledge have shown very strong supervised performance on commonsense reasoning, the solution process is usually complex and expensive. Generally, we should gather task-specific training data and then train models to learn the patterns in data. However, as shown in WinoGrande (Sakaguchi et al., 2020), acquiring unbiased labels requires a carefully designed crowd-sourcing procedure, which greatly adds to the cost of data collection. Moreover, supervising on large training sets is usually time-consuming. Therefore, instead of applying specific methods to the corresponding task, a reasonable framework is to convert diverse commonsense reasoning tasks to a common task and use a general unsupervised method to solve it. Furthermore, some tasks that lack sufficient annotations can be solved by the framework.

We attempt to use Natural Language Inference (NLI) as the common task mentioned above. NLI is the task of determining whether a hypothesis is “entailment” or “not entailment” to a given premise.

*Corresponding Author

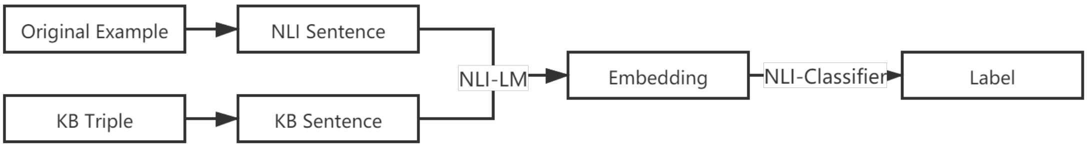

Figure 1: Overview of our NLI framework with injected knowledge from the knowledge base. The NLI-LM and NLI-Classifier denote LM with a classification head fine-tuned on NLI. We convert the original example of the source task, e.g., WinoWhy or CommonsenseQA, to NLI form and combine KB sentences as input.

NLI task is well-suited to be a common task, as it assembles the skills involved in sentence understanding, from the resolution of syntactic ambiguity to pragmatic reasoning with world knowledge (Wang et al., 2019b). Furthermore, the NLI task has been actively studied, especially since the emergence of large-scale datasets (Bowman et al., 2015; Williams et al., 2018), and we can directly leverage the progress. Moreover, we explore whether injecting external knowledge from knowledge bases to our framework can enhance the model's performance over commonsense reasoning tasks.

We apply RoBERTa (Liu et al., 2019), which has shown powerful performance on NLI tasks, as the backbone network of our NLI framework. We first convert a commonsense reasoning task to an NLI form as the original input. As shown in Table 1, we replace the pronoun "she" of the original WSC sentence with the correct candidate "Joan" and treat the replaced sentence as the hypothesis, while the given reason is the premise. Next, to leverage external knowledge, we use the recently-introduced ATOMIC 2020 (Hwang et al., 2020) and ConceptNet as knowledge bases (KBs). Specifically, we extract KB triples from KB by matching semantic similarity between the embeddings of KB and the source task and then combine the triples and the original input for RoBERTa. Our experimental results on WinoWhy and CommonsenseQA suggest that the NLI framework is suitable for commonsense tasks and external knowledge can provide useful information to help the model make the correct prediction. Furthermore, more improvements can be obtained by combining multiple effective categories of knowledge. In addition, models perform worse when facing problems about quantity knowledge and antonym relation.

2 Method

In this section, we describe the details of 1) using the NLI framework to solve commonsense reasoning tasks and 2) extracting knowledge from KBs, and 3) injecting the external knowledge into the NLI framework. The overview of our framework is shown in Figure 1.

2.1 NLI Task: A General Framework

The key to solving commonsense reasoning tasks such as WinoWhy and CommonsenseQA is to determine the relation between question-answer pairs. Follow this intuition, we use a general task NLI, which aims at identifying whether a hypothesis sentence can be entailed by a premise sentence. We first convert the original example of the source task to the NLI form. In this work, we define the source task as to predict whether an answer is entailed given a question. We can find that the question corresponds to the premise and the answer to the hypothesis. Moreover, for source tasks like WinoWhy, we can also try to convert the question (e.g., WSC question shown in Table 1) to a statement as the hypothesis, and treat the reason as a premise, following the if-then relation. Then, we use pre-trained language models (LM) with a classification head to solve the NLI task. Specifically, given a premise and a hypothesis, we concatenate them as the "NLI sentence": `[CLS]` Premise `[SEP]` Hypothesis `[SEP]`. The LM with the classification head then predicts the entailment relation.

To mitigate the data scarcity in an unsupervised setting, we consider transferring knowledge from large NLI datasets. Specifically, we fine-tuned the LM and classification head on either MNLI (Williams et al., 2018) or QNLI (Wang et al., 2019b). We use the RobertaForSequenceClassification from the transformers library (Wolf et al., 2020). It is the RoBERTa with a classification head on top. When evaluating our framework on source task, because MNLI has three labels: "entailment", "neutral", and "contradiction", we treat the last two labels as "not entailment". We denote the LM and classification head fine-tuned on NLI datasets as

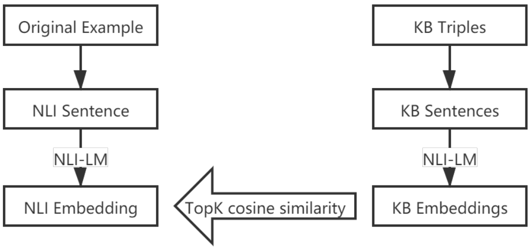

Figure 2: The method of extracting knowledge from KB. The NLI sentence is composed of the form “`[CLS]` Premise `[SEP]` Hypothesis `[SEP]`” and the KB sentence is wrapped by `[CLS]` and `[SEP]` as well. The NLI-LM denotes LM fine-tuned on NLI datasets.

NLI-LM and NLI-classifier.

2.2 Inject External Knowledge

In the following, we show how to extract knowledge from KB and inject the matched knowledge into the NLI framework. We first convert the triples in KB to natural language sentences and extract KB triple from KB by calculating cosine similarity between the embeddings of KB sentence and source task example. Finally, we combine the external KB sentence and original example to help NLI-LM and NLI-Classifier perform the correct prediction.

Convert KB Triple to Natural Language Inspired by ATOMIC (Sap et al., 2019a), which is unique in that the entity in a triple is mostly short sentences, we try to convert KB triple to natural language sentence, then capture helpful knowledge for original example by matching semantic similarity between them. For example, (*PersonX thanks PersonY afterwards*, isAfter, *PersonX asked PersonY for help on her homework*), a triple in ATOMIC, can be extended to “*After PersonX asked PersonY for help on her homework, PersonX thanks PersonY afterwards*”, and (*having_no_food*, CausesDesire, *go_to_a_store*), a triple in ConceptNet, corresponds to “*having no food makes someone want go to a store*”. In this work, we use ATOMIC 2020 (we call it ATOMIC for short in the following) and ConceptNet as knowledge bases. We define templates for every relation in ATOMIC and ConceptNet. Then we convert the triples in ATOMIC and ConceptNet to natural language sentences automatically using the templates. We named the natural language sentence “KB sentence”.

Extract Knowledge from KB Inspired by (Reimers and Gurevych, 2019), we use NLI-LM to generate the token embeddings of an NLI sentence or a KB sentence. Then we compute the mean of all token embeddings. As shown in Figure 2, an NLI sentence and a KB sentence are input into NLI-LM, and two mean embeddings are output. Then we calculate the cosine similarity between two embeddings as the semantic similarity of two input sentences. When input into NLI-LM, a NLI sentence is composed of the form “`[CLS]` Premise `[SEP]` Hypothesis `[SEP]`” and a KB sentence is wrapped by `[CLS]` and `[SEP]` as well. When evaluating our framework on a commonsense dataset, for each example, we extract the KB sentences with TopK semantic similarity.

Inject KB Sentence into NLI Sentence To combine a KB sentence and an NLI sentence, we inject the KB sentence into the middle of the NLI sentence to form a combined sentence. Thus, the form of the combined sentence is “`[CLS]` Premise `[SEP]` KB sentence `[SEP]` Hypothesis `[SEP]`”. For an NLI sentence with TopK matched KB sentences, we can generate K combine sentences. All of them are input into NLI-LM and K `CLS`-token embeddings are output. Then we compute the mean of all `CLS`-token embeddings. Finally, the mean embedding is input into the NLI-Classifier and the entailment relation is output.

3 Experiments

3.1 Tasks

We evaluate our framework on two commonsense reasoning datasets, WinoWhy and CommonsenseQA. Both require commonsense knowledge beyond textual understanding to perform well. All of our experiments use an unsupervised setting, i.e., our model does not train on the source task.

WinoWhy (Zhang et al., 2020a) contains 2,865 reasons, which belongs to 273 WSC examples respectively. We evaluate models on the full set. A WinoWhy example consists of a WSC question and reason. We use two strategies to convert an example to an NLI sentence. As the example shown in Table 1, (a) we directly treat the WSC question as premise and reason as a hypothesis. Then the NLI sentence is “`[CLS]` WSC question `[SEP]` reason `[SEP]`”. (b) we replace the asked pronoun in the WSC sentence with the correct candidate and treat the replaced sentence as a hypothesis, while the given reason now is the premise. Then the NLI sentence is "[CLS] reason [SEP] replaced WSC sentence [SEP]". According to the experimental results, LM fine-tuned on MNLI uses strategy (a), while LM fine-tuned on QNLI use strategy (b).

CommonsenseQA is a multiple-choice QA dataset that specifically measures commonsense reasoning. This dataset is constructed based on ConceptNet. We evaluate models on the development set with 1,221 questions since the answers to the test set are not publicly available. A CommonsenseQA example consists of a question and 5 choices. We regard the question and a choice as a NLI sentence with the form "[CLS] Q: question [SEP] A: choice [SEP]" (The additional "Q" and "A" follows the recommendation from the FairSeq repo on how to fine-tune RoBERTa on CommonsenseQA[1]). Then entailment score of every choice is calculated. Finally, the choice with the highest score is selected as the answer to the question. In addition, the form of combine sentence is "[CLS] Q: question [SEP] K: ATOMIC sentence [SEP] A: choice [SEP]".

3.2 Knowledge Bases

ATOMIC (Sap et al., 2019a) is a knowledge base consists of 880K of triples across 9 relations that cover social commonsense knowledge, e.g., (X gets X's car repaired, xIntent, to maintain the car), including aspects of events such as mental states, personal attribute, and social effect. As the later work, (Hwang et al., 2020) extends ATOMIC to ATOMIC 2020 with 1.33M triples. ATOMIC 2020 introduces 23 commonsense relations. Triples are of the form ({Event | Entity}, r, {Entity | Event | Behavior | Persona | Mentalstate}), where head and tail are nouns or short sentences and r represents an if-then relation type or physical property (e.g., xIntent and ObjectUse). We define 23 templates for every relation in ATOMIC 2020 to automatically convert triple to natural language sentences.

ConceptNet (Speer et al., 2017) is a knowledge base focus mostly on taxonomic and lexical knowledge (e.g., IsA, PartOf) and physical commonsense knowledge (e.g., MadeOf, UsedFor). We extracted 29 relations to form a subset with 485K entity-relation triples. Similar to ATOMIC, we define 29 templates for every relation. In this work, we use ATOMIC 2020 and ConceptNet for injecting external knowledge to NLI framework.

Models	**RoBERTa-Base** Full-Acc.(%)	**RoBERTa-Large** Full-Acc.(%)
Random	50.00	50.00
Original	55.78	55.67
+WinoGrande	56.19	58.18
+MNLI	66.87	70.61
+QNLI	70.40	70.86
+MNLI+CN	66.70	70.92
+QNLI+CN	72.46	71.10
+MNLI+ATOMIC	67.23	71.13
+QNLI+ATOMIC	**72.81**	**73.47**

Table 2: Performance comparison on the full set of WinoWhy. "Original" denotes the original LM. "+WinoGrande/MNLI/QNLI" denotes LM fine-tuned on these datasets. "+CN/ATOMIC" denotes LM with knowledge either from ConceptNet or ATOMIC. The accuracies of Original and +WinoGrande are reported by (Zhang et al., 2020a), while accuracies below are achieved by our framework.

Models	Dev-Acc.(%)	Dev-Acc.(%)
Random	20.00	-
Self-Talk	32.40	-
SMLM	38.80	-
BERT-Base Sup.	52.60	-
	RoBERTa-Base	RoBERTa-Large
Original	19.98	20.48
+MNLI	27.52	29.73
+QNLI	37.02	35.87
+MNLI+CN	31.70	29.24
+QNLI+CN	39.89	48.98
+MNLI+ATOMIC	31.70	29.16
+QNLI+ATOMIC	**42.10**	**52.09**

Table 3: Performance comparison on the dev set of CommonsenseQA. The accuracies of Self-Talk and SMLM are reported by (Shwartz et al., 2020) and (Banerjee and Baral, 2020). "BERT-Base Sup." denote the base model of BERT training on CommonsenseQA training set and the result is the accuracy of the test set reported by the official leaderboard.

3.3 Baselines

For WinoWhy, we consider the pre-trained language model: the base and large model of RoBERTa, as they show promising results on WSC. RoBERTa is a recently improved version of BERT (Devlin et al., 2019) with a larger amount of training instances and techniques such as dynamic masking, which performs consistently better than BERT over many benchmark datasets. A later work (Sakaguchi et al., 2020) has further enhanced the performance by fine-tuning RoBERTa with a larger and more balanced dataset WinoGrande. In our experiments, we denote the base and large model as RoBERTa-Base and RoBERTa-Large respectively. And we denote LM fine-tuned on WinoGrande as

[1] https://github.com/pytorch/fairseq/tree/master/examples/roberta/commonsense_qa

Models	WinoWhy Full-Acc.(%)	CommonsenseQA Dev-Acc.(%)
Overall	73.47	52.09
+Physical-Entity	67.36	48.98
+Event-Centered	73.08	51.42
+Mental-State	72.64	50.35
+Persona	71.69	50.20
+Behavior	73.02	51.09
+Behavior &Event-Centered	73.57	53.15

Table 4: Effect of ATOMIC category on WinoWhy and CommonsenseQA. We divide ATOMIC into five categories and inject each category separately into RoBERTa-Large + QNLI. Then we combine the two most effective categories.

+WinoGrande. We directly use the results reported by (Zhang et al., 2020a).

Same as WinoWhy, we use RoBERTa as baselines for CommonsenseQA. Specifically, we use RobertaForMaskedLM from the transformers library (Wolf et al., 2020). It can be regarded as a RoBERTa Model with a masked language modeling head on top. Given a CommonsenseQA question and one of the five choices, we mask the choice tokens and use the masked LM head to predict them. For example, a CommonsenseQA sentence input to model consists of the form: "`[CLS]` question `[SEP]` choice `[SEP]`". Then the choice will be masked and the masked LM head is used to predict the cross-entropy loss for it. Finally, the choice with the lowest loss will be selected as the answer to the question.

In addition, we compare our model with Self-Talk (Shwartz et al., 2020) and SMLM (Banerjee and Baral, 2020). These two models both propose an unsupervised framework to multiple-choice commonsense tasks and show considerable improvements over large pre-trained language models. So we report their dev-set accuracies on CommonsenseQA as baselines.

4 Results and Analysis

4.1 Main Results

Table 2 and Table 3 show results of applying NLI framework and external knowledge to WinoWhy and CommonsenseQA. Our framework has achieved state-of-the-art (SOTA) unsupervised performance on WinoWhy by a large margin. Specifically, using the same language model RoBERTa, we observed improvements ranging from +8.52% (66.70% by Base+MNLI+CN) to +15.29% (73.47% by Large+QNLI+ATOMIC) compared to the previous SOTA result (58.18%).

Models	WinoWhy Full-Acc.(%)	CommonsenseQA Dev-Acc.(%)
Overall	71.10	48.98
+Physical-Entity	67.47	49.88
+Event-Centered	71.24	50.94
+Social-Interaction	66.49	47.83
+Taxonomic-Lexical	71.24	46.52
+Physical-Entity &Event-Centered	69.81	51.76
+Taxonomic-Lexical &Event-Centered	71.58	49.96

Table 5: Effect of ConceptNet category on WinoWhy and CommonsenseQA. We divide ConceptNet into four categories and inject each category separately into RoBERTa-Large + QNLI. Then we combine the two most effective categories on WinoWhy and CommonsenseQA, respectively.

As for results on CommonsenseQA, we first observe that RoBERTa is struggling near the Random Guess baseline. This result illustrates that RoBERTa completely cannot deal with CommonsenseQA without training. However, after converting CommonsenseQA to NLI form and injecting KB sentences, RoBERTa behaves a lot better. RoBERTa-Base + MNLI + CN/ATOMIC gets a comparable result compared to Self-Talk, while RoBERTa-Base + QNLI + CN/ATOMIC have already exceeded SMLM, the previous SOTA method. Finally, it is interesting to note that RoBERTa-Large + QNLI + ATOMIC is slightly worse than BERT-Base model training on the CommonsenseQA training set.

Now we focus on results applying the NLI framework without injected knowledge. For WinoWhy, RoBERTa can achieve a considerable boost after being fine-tuned on either QNLI or MNLI. For CommonsenseQA, RoBERTa fine-tuned on QNLI can get +4.62% higher dev-set accuracy than Self-Talk and comparable result to SMLM. The experiment clearly illustrates the effectiveness of the NLI framework and transfer learning from NLI datasets.

When we inject KB sentences to RoBERTa fine-tuned on QNLI, improvement can be observed on full-set accuracy for WinoWhy and dev-set accuracy for CommonsenseQA. This indicates that the knowledge from QNLI and that extracted from KB complement each other. On the other hand, external knowledge, either from ATOMIC or ConceptNet, is not much help to RoBERTa fine-tuned on MNLI and even causes a drag. We hypothesize that there

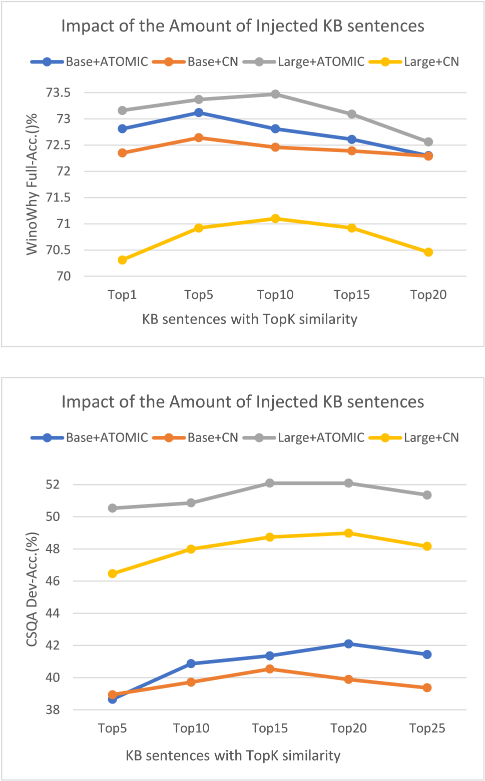

Figure 3: Effect of K on WinoWhy and CommonsenseQA. We inject KB sentences with topK similarity into NLI-LM. The figure above is the results on WinoWhy, while the below is on CommonsenseQA. The "Base" and "Large" denote RoBERTa-Base/Large + QNLI.

is a high overlap or even contradiction between the knowledge of KB and MNLI, which causes the incompatibility between them.

In summary, We think the reasons leading to the significant improvement are 1) NLI framework is better suited for such tasks; 2) RoBERTa picks up necessary knowledge from the NLI datasets; 3) ATOMIC and ConceptNet provide some useful information to source tasks and help models make the correct prediction.

4.2 Ablation Study

Category of Injected KB Sentences In order to study whether the different categories of external knowledge will have a large impact on the model's performance, we divide ATOMIC into five categories: Physical-Entity, Event-Centered, Mental-State, Persona, and Behavior, basically following the definition of (Hwang et al., 2020). Physical-Entity deals with inferential knowledge about common entities and objects. Event-Centered provides intuitions about how common events are related to one another. Mental-State addresses the emotional or cognitive states of the participants in a given event. Persona describes a person's attribute as perceived by others given an event. Behavior address the socially relevant responses to an event. We inject each category separately into RoBERTa-Large + QNLI (the best NLI-LM in our experiment). The results are shown in Table 4. Similar to ATOMIC, we divide ConceptNet into four categories: Physical-Entity, Event-Centered, Social-Interaction, and Taxonomic-Lexical. The meanings of the former two categories are the same as ATOMIC. Social-Interaction focuses on socially triggered states and behaviors. Taxonomic-Lexical focus on taxonomic and lexical. The results are shown in Table 5. It is not surprising that there are some categories of knowledge dragging down the performance. For example, for ATOMIC, injecting the knowledge of Physical-Entity obtains the worst results on either WinoWhy or CommonsenseQA.

Next, we wonder if we can get higher accuracies after combining the effective categories. So we combine the two most effective categories for each task. On ATOMIC, they are Behavior and Event-Centered. On ConceptNet, they are Taxonomic-Lexical and Event-Centered for WinoWhy, while Physical-Entity and Event-Centered for CommonsenseQA. The results show that this strategy makes the performance exceed the "Overall". The slight boost of accuracies illustrates that our assumption is correct. We also find that combining any two categories does not necessarily work through the results of ConceptNet. For example, combining Taxonomic-Lexical and Event-Centered does not get a higher result than "Overall" on CommonsenseQA, because of the bad performance of Taxonomic-Lexical. It tells us that we need to identify effective knowledge when combining different categories.

Amount of Injected KB Sentences As mentioned before, we extract KB sentences with topK similarity. Now we investigate the impact of hyperparameter K with experimental results shown in Figure 3. The results generally follow the intuition that the more knowledge is injected, the better the performance is until the amount of injected sentence reaches a threshold. Then the accuracy begins to decrease. We think the reason is that knowledge with lower semantic similarity introduces noise to the model and then plays a distracting effect.

Models	Property	Object	Eventuality	Spatial	Quantity
	(337)	(856)	(928)	(676)	(206)
RoBERTa-Large					
+WinoGrande	56.08	58.06	59.59	56.82	56.80
RoBERTa-Base					
+QNLI	72.40	68.81	71.01	70.33	66.50
+QNLI+CN	73.89(+1.49)	72.20(+3.39)	72.09(+1.08)	73.00(+2.67)	67.96(+1.46)
+QNLI+ATOMIC	75.37(+2.97)	73.01(+4.20)	73.17(+2.16)	74.04(+3.71)	67.69(+1.19)
RoBERTa-Large					
+QNLI	73.89	69.98	70.12	72.26	70.87
+QNLI+CN	74.69(+0.80)	70.86(+0.88)	70.13(+0.01)	72.47(+0.21)	69.87(-1.00)
+QNLI+ATOMIC	76.56(+2.67)	71.26(+1.28)	72.95(+0.83)	74.18(+1.92)	70.87(+0.00)

Table 6: Performance comparison on different knowledge type set of WinoWhy. WSC questions are grouped by their major knowledge types. If one question contains more than one knowledge type, it will be counted in all types. The numbers of examples are shown in brackets. The results of RoBERTa-Large + WinoGrande are reported by (Zhang et al., 2020a).

Models	AtLocation	Causes	CapableOf	Antonym	HasPrerequisite
	(526)	(175)	(101)	(72)	(43)
RoBERTa-Base					
+QNLI	35.36	42.43	41.58	29.17	34.88
+QNLI+CN	39.16(+3.80)	42.86(+0.43)	42.62(+1.04)	29.33(+0.16)	37.21(+2.33)
+QNLI+ATOMIC	41.83(+6.47)	48.00(+5.57)	44.59(+3.01)	31.33(+2.16)	39.53(+4.65)
RoBERTa-Large					
+QNLI	34.41	42.29	39.60	34.72	37.21
+QNLI+CN	46.01(+11.60)	53.71(+11.42)	50.50(+10.90)	41.67(+6.95)	65.12(+27.91)
+QNLI+ATOMIC	51.14(+16.73)	56.00(+13.71)	52.48(+12.88)	41.67(+6.95)	69.77(+32.56)

Table 7: Performance comparison on different knowledge type set of CommonsenseQA. Questions are classified based on the ConceptNet relation between the question concept and correct answer concept. We select the relations that have more than 40 questions as knowledge types. The numbers of examples are shown in brackets.

4.3 Discussion

To discuss the performance when the model faces different knowledge types, we follow the knowledge types defined in (Zhang et al., 2020a) and divide WinoWhy into five subsets. We evaluate RoBERTa-Base/Large + QNLI on each subset. The results are shown in Table 6. "Property" denotes the knowledge about the property of objects. "Object" represents that about objects. "Eventuality", "Spatial" and "Quantity" corresponding to eventualities, spatial position, and numbers, respectively. Comparing RoBERTa-Large fine-tuned on WinoGrande (the best model reported by Zhang et al., 2020a) and RoBERTa fine-tuned on QNLI, the latter goes beyond the former on all knowledge types. It is no doubt that QNLI contains more commonsense knowledge needed by WinoWhy than WinoGrande. Now let us focus on the comparison between models with and without KB sentence. It is shown that KB sentences matched for WinoWhy examples can provide some performance boost on most knowledge types, suggesting that we successfully inject the effective knowledge from KB to RoBERTa. Further, we can find that whether for RoBERTa + QNLI or RoBERTa + QNLI + KB, the worst performances appear on "Quantity". What's more, the injected knowledge, either from ATOMIC or ConceptNet, brings the lowest benefit to "Quantity", and even results in the only drag for RoBERTa-Large + QNLI + CN (-1.00). The reason for this result may be due to the lack of knowledge about numbers in QNLI, ATOMIC, and ConceptNet. We can find that large corpora do often lack quantity knowledge. This gives us the idea that constructing and encoding quantity knowledge into LM in the future.

Similar to WinoWhy, we follow the experiment described in (Ma et al., 2019) and divide CommonsenseQA into five subsets. We classify questions based on the ConceptNet relation between the question concept and the correct answer concept. Then we select the relations with more than 40 questions as knowledge types. Observing experimental results shown in Table 7, we can derive the same conclusion as WinoWhy that injecting knowledge following our method can provide useful information to LM and help make the correct decision. However, accuracies on "Antonym" are the lowest compared with other knowledge types. And the boosts are also the lowest after injecting knowledge. "Antonym" denotes that A and B are opposites in some relevant way, such as black and white. We guess it is because the language model has a weak ability to deal with antonym relations.

In addition, we find that ATOMIC can bring more benefits to RoBERTa compared with ConceptNet. As described in (Hwang et al., 2020), triples in ConceptNet are limited to mostly taxonomic, lexical, and object-centric physical knowledge, making the commonsense portion of ConceptNet relatively small. While ATOMIC has more knowledge related to social commonsense, and relatively, the coverage is more extensive and balanced. Our experimental results are consistent with these descriptions.

5 Related work

Commonsense Reasoning Recent commonsense reasoning datasets (Bhagavatula et al., 2020; Zhou et al., 2019; Sap et al., 2019b; Bisk et al., 2020; Talmor et al., 2019) have motivated research in several domains of commonsense: abductive, temporal, social, and physical. SOTAs for most of them have achieved over 80% accuracy, which is close to human performance (e.g., Brown et al., 2020; Khashabi et al., 2020; Raffel et al., 2020). However, their success is due to larger pre-trained corpora and much more parameters, which is difficult to be followed for most researchers. In addition, other useful methods (Yasunaga et al., 2021; Feng et al., 2020; Wang et al., 2020) generally require training on training sets and knowledge graphs. When applying them to different tasks, the same running and tuning process should repeat for several times to find the best fit. Thus, we propose a framework to convert diverse commonsense reasoning tasks to a common task, NLI, and use a general unsupervised method to solve it.

Natural Language Inference Since GLUE regards NLI as a benchmark task for testing the natural language understanding capability of the model, NLI has been well studied, and language models have achieved performance beyond humans on some NLI datasets. Furthermore, by leveraging transfer learning from large NLI datasets, great performances have been achieved in several tasks, such as story ending prediction (Li et al., 2019), intent detection (Zhang et al., 2020b), semantic textual similarity (Reimers and Gurevych, 2019). Therefore, we attempt to use NLI as the common task to solve commonsense reasoning.

External Knowledge Most commonsense reasoning tasks require models to synthesize external commonsense knowledge and leverage more sophisticated reasoning mechanisms. The key is to extract effective information from commonsense sources, such as ATOMIC, ConceptNet, and Wikipedia. Methods learn commonsense knowledge either by KGs pre-training (Bosselut et al., 2019; Bosselut and Choi, 2019; Ye et al., 2019) or by reasoning on knowledge graphs (Feng et al., 2020; Lv et al., 2020; Lin et al., 2019). In order to cooperate with our NLI framework, we convert the triples in KB to natural language sentences and extract triples by calculating cosine similarity between the embeddings of KB sentence and source task example.

6 Conclusion

In this work, we propose a framework to convert diverse commonsense reasoning tasks to a common task, NLI and use a pre-trained language model, RoBERTa to solve it. By leveraging transfer learning from large NLI datasets, QNLI and MNLI, and injecting crucial knowledge from knowledge bases such as ATOMIC and ConceptNet, our framework achieved SOTA unsupervised performance on two commonsense reasoning tasks: WinoWhy and CommonsenseQA. Experimental results show that knowledge from QNLI and extracted from either ATOMIC or ConceptNet can complement each other to enhance the model's performance on commonsense reasoning. More improvements can be obtained by combining multi categories of effective knowledge. Further experiment shows that ATOMIC can bring more benefits to RoBERTa compared with ConceptNet. However, injected knowledge is not much help to RoBERTa fine-tuned on MNLI and even causes a drag. In addition, models perform worse when facing problems about quantity knowledge and antonym relation. The code is publicly available: `https://github.com/sysuhcm/NLI-KB`.

Acknowledgement

We acknowledge support from the Natural Science Foundation of China under Grant No. 62076261.

References

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 151–162. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering. *CoRR*, abs/1911.03876.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ernest Davis. 2017. Logical formalizations of commonsense reasoning: A survey. *J. Artif. Intell. Res.*, 59:651–723.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1295–1309. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. *CoRR*, abs/2010.05953.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 1896–1907. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. AAAI Press.

Zhongyang Li, Xiao Ding, and Ting Liu. 2019. Story ending prediction by transferable BERT. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1800–1806. ijcai.org.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2829–2839. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8449–8456. AAAI Press.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. *CoRR*, abs/1910.14087.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

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 2020, Online, November 16-20, 2020*, pages 4615–4629. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro A. Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4129–4140. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. *CoRR*, abs/2104.06378.

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. *CoRR*, abs/1908.06725.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020a. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5736–5745. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip S. Yu, Richard Socher, and Caiming Xiong. 2020b. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5064–5082. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3361–3367. Association for Computational Linguistics.