Leveraging Knowledge in Multilingual Commonsense Reasoning

Yuwei Fang, Shuohang Wang, Yichong Xu, Ruochen Xu, Siqi Sun
Chenguang Zhu, Michael Zeng
Microsoft Cognitive Services Research Group
{yuwfan, shuowa, yicxu, ruox, siqi.sun, chezhu, nzeng}@microsoft.com

Abstract

Commonsense reasoning (CSR) requires models to be equipped with general world knowledge. While CSR is a language-agnostic process, most comprehensive knowledge sources are restricted to a small number of languages, especially English. Thus, it remains unclear how to effectively conduct multilingual commonsense reasoning (XCSR) for various languages. In this work, we propose to use English as a pivot language, utilizing English knowledge sources for our our commonsense reasoning framework via a translate-retrieve-translate (TRT) strategy. For multilingual commonsense questions and answer candidates, we collect related knowledge via translation and retrieval from the knowledge in the source language. The retrieved knowledge is then translated into the target language and integrated into a pre-trained multilingual language model via visible knowledge attention. Then we utilize a diverse of four English knowledge sources to provide more comprehensive coverage of knowledge in different formats. Extensive results on the XCSR benchmark demonstrate that TRT with external knowledge can significantly improve multilingual commonsense reasoning in both zero-shot and translate-train settings, consistently outperforming the state-of-the-art by more than 3% on the multilingual commonsense reasoning benchmark X-CSQA and X-CODAH.

1 Introduction

Commonsense reasoning (CSR) is one of the key challenges in natural language understanding. It requires a model to integrate world knowledge into language modeling to produce answers. A large number of tasks have been proposed to evaluate commonsense reasoning in English, such as COPA (Roemmele et al., 2011a) and CSQA (Talmor et al., 2019).

Most recently, multilingual commonsense reasoning (XCSR) extends a model's commonsense capability beyond language barriers and has begun to draw attention from the community. A number of multilingual datasets have emerged for this challenging task, for example, X-CSQA (Lin et al., 2021), X-CODAH (Lin et al., 2021), XCOPA (Ponti et al., 2020).

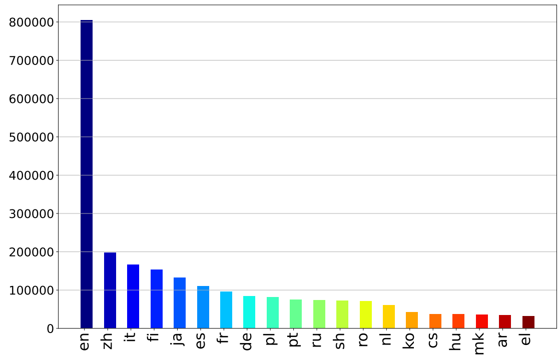

Figure 1: Number of total definitions per language. The statistics are generated from Wiktionary 2021-10-01 dump. There are 55 languages with 10,000 or more definitions and we list top 20 languages by the definition count here.

To solve commonsense reasoning tasks, it is essential to fuse human created knowledge into pre-trained language models (PLM) (Lin et al., 2019; Feng et al., 2020; Yu et al., 2020; Xu et al., 2021b). For example, DEKCOR (Xu et al., 2021b) integrates knowledge from ConceptNet (Speer et al., 2017) and Wiktionary [1] into the ALBERT model (Lan et al., 2020) for commonsense question answering. However, most existing knowledge sources are crafted in a few popular languages, especially English. For example, Figure 1 shows the number of total definitions in English is much greater than any other language based on the statistics from a recent 2021-10-01 dump of Wiktionary. Thus, it remains an open question how to tackle XCSR with a lack of curated knowledge in the

[1] https://www.wiktionary.org/

Findings of the Association for Computational Linguistics: ACL 2022, pages 3237 - 3246
May 22-27, 2022

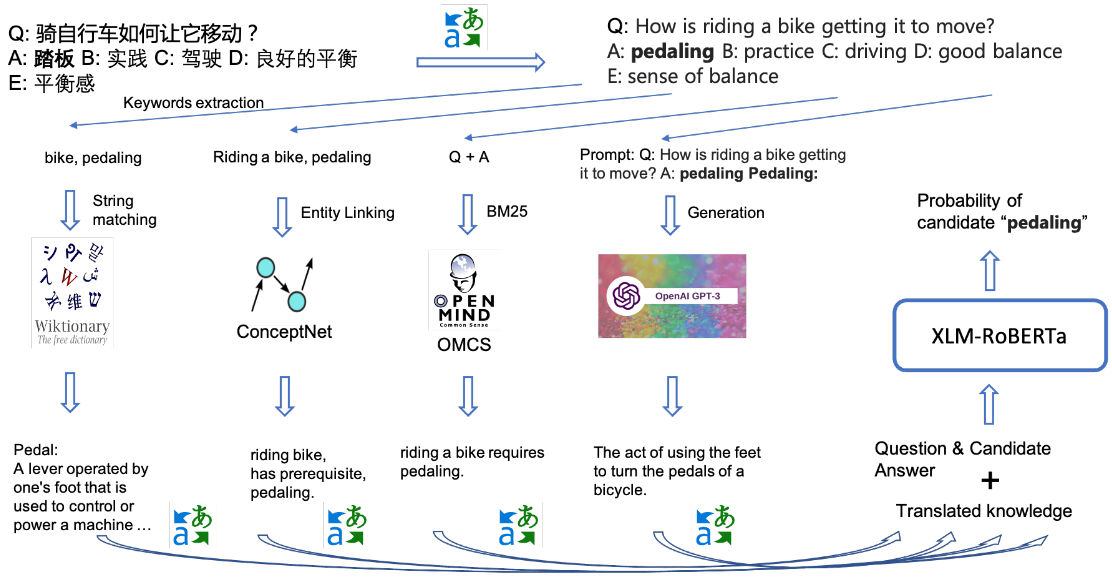

Figure 2: An overview of our framework for multilingual commonsense reasoning. Given the question and candidate answers in the target language (Chinese), we first translate it into English, then retrieve related knowledge from four English knowledge sources and translate the retrieved knowledge back into the target language. The retrieved knowledge, along with question and candidate answer, are fed into the multilingual pretrained language model for answer prediction.

target language.

In this paper, we propose a translate-retrieve-translate (TRT) solution to utilize English knowledge sources for XCSR. Specifically, given a commonsense reasoning question (possibly concatenated with a candidate answer) in the target language, we first translate it into English. Next, we retrieve related knowledge from English knowledge sources. The retrieved knowledge is then translated back into the target language. Finally, the knowledge is integrated into a multilingual language model via an visible knowledge attention mechanism to answer the question.

Another contribution of our work is that the utilization of a diverse set of knowledge sources to provide a more comprehensive coverage of knowledge in different formats. Specifically, we utilize unstructured text corpus (Open Mind Common Sense (Singh, 2002)), structural knowledge graph (ConceptNet (Speer et al., 2017)), dictionary (Wiktionary) and large-scale language model (GPT-3 (Brown et al., 2020)). Given an input query, we utilize information retrieval, entity linking, and model inference to obtain knowledge from the corresponding sources.

We conduct extensive evaluation of our model on the multilingual commonsense reasoning benchmark X-CSQA and X-CODAH (Lin et al., 2021). The results demonstrate the effectiveness of our proposed translate-retrieve-translate solution with multiple knowledge sources. For example, in the zero-shot transfer setting, TRT with Wiktionary can improve 1.9 and 2.7 points over the baselines. For translate-train setting, TRT with Wiktionary and OMCS outperform 1.6 and 1.0 over the baselines. Compared with previous state-of-the-art results on the XCSR leaderboard, TRT improve them by more than 3 points.

We summarize the main contributions of this work as follows. (i) We propose a translate-retrieve-translate (TRT) solution to utilize English knowledge sources for multilingual commonsense reasoning. (ii) We comprehensively explore four knowledge sources in different formats and demonstrate their utility on a pair of XCSR benchmarks (X-CSQA, X-CODAH). (iii) We achieve the state-of-the-art results on the on XCSR leaderboard, outperforming the previous best methods by more than 3.3 points.

2 Related Work

Multilingual Commonsense Reasoning Evaluating the commonsense reasoning abilities of trained models has been explored through a variety of tasks and problem settings. An early work in this space was the Winograd scheme challenge (Levesque et al., 2012), where the goal is to disambiguate the reference of a pronoun (Levesque

Knowledge Source	Knowledge Format	Query Format	Retrieved Knowledge	Retrieval Method
Wiktionary	Dictionary	Content Word	Definition	String Matching
ConceptNet	Entity-Relation Triplets	Entity Pair	Entity-Relation Triplet	Entity linking
OMCS	Text in Sentences	Sentences	Sentences	BM25
GPT-3	Parameters	Unstructured Text	Unstructured Text	Conditional Generation

Table 1: Different knowledge resources for retrieval. OMCS is short for Open Mind Common Sense (Singh, 2002). Wiktionary is a multilingual, web-based dictionary from https://www.wiktionary.org/. ConceptNet (Speer et al., 2017) is a freely-available multlingual knowledge graph. GPT-3 (Brown et al., 2020) is a large scale pre-trained language model.

et al., 2012). Another early work is COPA (Roemmele et al., 2011b), where the goal is to select cause or result for a premise. Later on, researchers have constructed larger scale datasets, such as SWAG (?), CODAH (Chen et al., 2019), and CommonsenseQA (Talmor et al., 2019), for commonsense knowledge learning. Recently, commonsense reasoning tasks have been extended to multilingual setting, such as X-CSQA (Lin et al., 2021), X-CODAH (Lin et al., 2021), XCOPA (Ponti et al., 2020). In paper, we focus on training models to learn commonsense knowledge in multiple languages.

External Knowledge Fusion Knowledge bases are an important external data source to help models learn the ability of commonsense reasoning. A wide range of knowledge resources, such as ConceptNet (Speer et al., 2017), Wikipedia, Freebase (Pellissier Tanon et al., 2016), and some KBs in domain (Fader et al., 2011), can be fused into the model. LoBue and Yates (2011) explored how commonsense knowledge involved in recognizing textual entailments. Guan et al. (2020) utilize commonsense knowledge to generate reasonable stories. Bi et al. (2019) incorporate external Knowledge into question answering. Xu et al. (2021b) fuse the ConceptNet (Speer et al., 2017) and Wikionary into the model for solving CommonsenseQA. In this paper, we will follow this direction and explore how to leverage different knowledge sources for multiligual commonsense reasoning.

Multilingual Language Model Large scale multilingual pretrained language models (MPLM) (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020) have always been the most important backbone for solving multilingual tasks including commensense reasoning tasks. Knowledge bases have also been integrated into the pretraining process (Kassner et al., 2021a; Jiang et al., 2021). As shown in (Kassner et al., 2021a), there exist multilingual knowledge base. But they still lack the components or contexts to explicitly integrate knowledge and commonsense. And Lin et al. (2021) builds a commonsense probing dataset to improve the pre-trained MPLM for commonsense reasoning beyond English. Our work is orthogonal to these pre-trained methods and focus on fusing knowledge during finetuning.

GPT-3 Prompt learning Large scale pretrained language models like GPT-3 (Brown et al., 2020) have shown tremendous success on few-shot learning. There exists a large body of work on prompting to leverage the implicit knowledge from it (Li and Liang, 2021; Liu et al., 2021; Wei et al., 2021). In this work, we focus on leveraging GPT-3 to generate three diverse knowledge formats and then fusing them into fine-tuning stage.

3 Approach

In this section, we first formalize the multilingual commonsense reasoning (XCSR) task (Section 3.1). Then we describe more details about our commonsense knowledge resources (Section 3.2). Next, we introduce our proposed translate-retrieve-translate (TRT) solution to obtain the multilingual knowledge (Section 3.3). Finally, we introduce how to fuse the obtained knowledge into multilingual pretrained language models by employing the visible attention mechanism (Section 3.4). An overview of our framework is illustrated in Figure 2.

3.1 Problem Formulation

We denote a language by $l \in L$, where $L = \{en, fr, de, zh, \cdots\}$. Given a commonsense question q^l in the target language l, the goal is to choose the correct answer from N candidates $\{c_1^l, c_2^l, \cdots, c_N^l\}$. We assume there are one or more external knowledge sources to provide world knowledge in various formats for commonsense reasoning. Each time the model retrieves

Dataset	Knowledge Source	Prompt
X-CODAH	Wiktionary	<Q> \n hedge: A thicket of bushes or other shrubbery, especially one planted as a fence between two portions of land.
	ConceptNet	<Q> \n hedge capable of fence house
	OMCS	<Q> \n he is a man.
X-CSQA	Wiktionary	<Q> \n pedalling: A lever operated by one's foot that is used to control or power a machine or mechanism, such as a bicycle or piano.
	ConceptNet	<Q> \n riding bike has prerequisite pedalling.
	OMCS	<Q> \n riding a bike requires pedalling.

Table 2: A GPT-3 prompt example with knowledge sources from Wiktionary, ConceptNet and OMCS. <Q> are short for the query "A man is using a pair of hedge trimmers on trees. He is talking to the camera as he goes." and "Q: How is riding a bike getting it to move? A: pedalling" for X-CODAH and X-CSQA datastes respectively.

knowledge from these sources using the question-candidate pair as query, i.e., $p^l = [q^l, c^l_i]$.

3.2 Commonsense Knowledge

External sources of commonsense knowledge are critical to the performance of a commonsense reasoning (CSR) model. Previous methods for CSR primarily integrate knowledge from one or two sources (Xu et al., 2021b). In this work, we conduct comprehensive experiments by leveraging commonsense knowledge from four different resources: unstructured text corpus (Open Mind Common Sense), knowledge graph (KG) (ConceptNet), dictionary (Wiktionary), and pre-trained language model (PLM) (GPT-3). Open Mind Common Sense (OMCS) (Singh, 2002) is a large commonsense knowledge base which has accumulated millions of facts from the contributions of many thousands of people across the Web. ConceptNet (Speer et al., 2017) is a freely-available semantic network, originated from OMCS. Wiktionary is a multilingual web-based project to create a free content dictionary and provides the definitions for all the words. GPT-3 (Brown et al., 2020) is a large-scale pre-trained language model which can be induced to generate knowledge for some queries (Liu et al., 2022). These knowledge resources are saved in quite diverse formats as the analysis shown in Table 1. To retrieve the knowledge, we will consider different query formats and retrieval methods in the next section.

3.3 Knowledge Retrieval

Most large-scale knowledge sources in either academia or industry are crafted in a few popular languages, especially in English (see Figure 1 as an example). To obtain knowledge for low-resource languages, we propose a translate-retrieve-translate (TRT) solution. In detail, we first use a machine translation tool to translate the query in all languages into English. Then, we can retrieve knowledge from English knowledge sources using the translated query. The retrieved knowledge can be then translated back into the original target language for model training.

As a knowledge source usually contains vast amounts of information, we need to retrieve and leverage only the related knowledge for a given query p^l. Next we introduce the details of knowledge retrieval for four knowledge sources.

Word definition retrieval from Wiktionary Every word has its own definition but not all of them are delivering knowledge for commonsense reasoning. In this work, we mainly focus on retrieving the content words, such as nouns, verbs, and adjectives, and the words harder to understand by multilingual language models. In detail, after part-of-speech tagging of the sequence, we select the nouns, verbs and adjectives as the candidate words. Then, we mask one word at a time and compute its masked language model (MLM) probability by pre-trained multilingual language model, XLM-RoBERTa (Conneau et al., 2020). We select top-N words with lowest MLM probability for dictionary retrieval. If the original word is not in Wiktionary, we try to find its lemmazied form. The first definition entry in Wiktionary is the retrieved knowledge.

Structured knowledge retrieval from ConceptNet A knowledge graph can provide relation information between entities. We enumerate pairs of candidate words from the input sequence and check whether there exists a relation between them in the knowledge graph ConceptNet. If so, we retrieve the corresponding triplet as the external knowledge.

Unstructured text retrieval from OMCS Open Mind Common Sense (OMCS) consists of knowledge in natural language description. We first build a search index [2] for all the sentences in OMCS. Then, whenever a new query comes, we retrieve the highest ranked sentence based on BM25 as the external knowledge text.

Knowledge Generation with GPT-3 Previous research shows that large-scale PLM contains rich knowledge implicitly (Roberts et al., 2020; Kassner et al., 2021b). Thus, we use one of the largest PLM, GPT-3 (Brown et al., 2020), to generate related knowledge given the query. As GPT-3 requires a prompt with input and output examples, we feed it with a few examples with a query and the knowledge in designated format. Table 2 lists an example with above three knowledge formats for X-CODAH and X-CSQA. For example, given the word 'pedalling'" and its definition "A lever operated by one's foot that is used to control or powera machine or mechanism, such as a bicycle or piano." along with the query " How is riding a bike getting it to move?", GPT-3 will generate its version of definition of a word it thinks important in the input query. For the prompt that is not in English, we translate the English prompt into the target language.

3.4 Fusing Knowledge into Multilingual Language Model

Given the question answer pair $p^l = [q^l, c_i^l]$, we use the retrieval techniques to collect K pieces of retrieved knowledge text: $S = [s_1, \cdots, s_K]$.

The most intuitive way is to concatenate them with p^l as input to the multilingual pre-trained language model (XPLM) for answer generation, i.e., the input would be $I =$ [CLS] q^l c_i^l [SEP] s_1 [SEP] $\cdots$ s_K [SEP].

However, this simple way may divert the original meaning of p^l because of the introduced noise by appending S, as pointed out by Liu et al. (2020); Xu et al. (2021a). To remedy this issue, we adopt the visibility matrix (Liu et al., 2020; Xu et al., 2021a) to limit the impact of knowledge set S on the original question-candidate pair p_l. Specifically, in each transformer layer of XPLM, an attention mask matrix M is added to the self-attention weights before softmax.

Suppose t_j and t_k are the j-th and k-th tokens from the input I. We set M_{jk} to zero to allow attention from t_j to t_k, and set M_{jk} to $-\infty$ to forbid attention. M_{jk} is set to zero if: i) both tokens belong to the input p_l, or ii) both tokens belong to the same knowledge s_i, or iii) t_j is the token at the start position of linked word in p_l and t_k is from its correspond knowledge text. More formally, the mask matrix M is

$$M_{jk} = \begin{cases} 0 & t_j, t_k \in p^l \\ 0 & t_j, t_k \in s_i \\ 0 & t_j \in p^l, t_k \in s_i \\ -\infty & \text{otherwise} \end{cases} \tag{1}$$

For model training, let $z_0 \in R^d$, the `[CLS]` hidden state from the last layer, denotes the representation of encoding the question, candidate, and the corresponding retrieved knowledge. d is the dimension of the output vector of the encoder. Then we calculate the prediction score $\hat{y}_i$ for each candidate c_i^l with one linear layer, $\hat{y}_i = W_o z_0$, where $W_o \in R^{1*d}$, followed by a softmax normalization upon all candidates, $\hat{y} = softmax([\hat{y}_i, \cdots, \hat{y}_N])$, where N is the number of candidate for each question. The final loss function is the standard cross-entropy loss.

Dataset	X-CSQA	X-CODAH
Task Format	QA	Scene Completion
#Languages	16	16
#Options	5	4
#train	8888	8476
#dev	1000	300
#test	1074	1000

Table 3: Statistics of the two datasets in the multilingual commonsense reasoning benchmark XCSR (Lin et al., 2021).

4 Experiments

In this section, we perform extensive experiments to explore the aforementioned TRT solution with four knowledge sources on the multilingual commonsense reasoning benchmark XCSR (Lin et al., 2021).

4.1 Datasets

Table 3 lists the statistics for the two datasets in XCSR. They are collected from CSQA (Talmor et al., 2019) and CODAH (Chen et al., 2019) by translating into another 15 languages other than English with online commercial services such as DeepL Pro Translate. (i) X-CSQA (Lin et al., 2021) for commonsense question answering: given

[2] https://lucene.apache.org/pylucene/

Dataset	Model	en	de	it	es	fr	nl	ru	vi	zh	hi	pl	ar	ja	pt	sw	ur	avg
X-CODAH	mBERT	42.9	33.1	33.5	33.8	35.2	33.7	31.9	22.8	38.0	26.5	31.0	34.8	34.0	37.2	30.8	31.5	33.2
	XLMR-B	50.1	45.8	44.4	44.2	45.2	42.0	44.1	43.2	44.6	38.1	41.9	37.8	42.0	44.1	35.6	34.6	42.4
	XLMR-L	66.4	59.6	59.9	60.9	60.1	59.3	56.3	57.4	57.3	49.1	57.5	51.2	53.8	58.2	42.2	46.6	56.0
	MCP (XLMR-L)	**69.9**	60.7	61.9	60.7	61.4	60.7	58.6	62.3	61.9	53.7	59.0	54.1	54.7	60.8	44.6	48.0	58.3
	TRT	69.1	**65.3**	**62.5**	**64.4**	**64.3**	**64.5**	**61.8**	**64.6**	**63.3**	**57.1**	**62.7**	**57.6**	**61.6**	**64.3**	**52.5**	**55.1**	**61.9**
X-CSQA	mBERT	38.8	29.6	36.4	35.3	33.8	32.6	32.7	22.2	37.8	21.1	27.2	27.7	31.4	34.1	21.8	23.7	30.4
	XLMR-B	51.5	44.1	42.1	44.8	44.0	43.3	39.5	42.6	40.6	34.6	40.2	38.4	37.5	43.4	29.6	33.0	40.6
	XLMR-L	66.7	56.1	58.2	59.5	60.3	56.8	52.1	51.4	52.7	48.7	53.9	48.4	50.0	59.9	41.6	45.2	53.8
	MCP (XLMR-L)	69.5	59.3	60.3	61.4	60.0	61.1	57.5	55.7	56.7	51.3	56.1	52.3	50.2	60.7	43.3	48.8	56.5
	TRT	**71.0**	**61.2**	**63.0**	**65.1**	**65.1**	**62.8**	**57.8**	**58.9**	**56.3**	**56.1**	**59.4**	**56.2**	**54.7**	**64.6**	**51.0**	**53.9**	**59.8**

Table 4: Overall test results on the multilingual commonsense reasoning benchmark XCSR. Results of mBERT (Devlin et al., 2019), XLMR-B, XLMR-L (Conneau et al., 2020), MCP(XLMR-L) (Lin et al., 2021) for X-CSQA and X-CODAH are from XCSR leaderboard (Lin et al., 2021). We submit the test prediction with the best dev result in table 5 to the XCSR leaderboard for evaluation. Leaderboard: https://inklab.usc.edu//XCSR/leaderboard

the human authored question that describes the relation between concepts from ConceptNet (Speer et al., 2017), the model needs to choose the answer from five concepts. All of the data in English are from original CSQA datset. (ii) X-CODAH (Lin et al., 2021) for Scene Completion: given a prompt question and the subject of the subsequence sentence, the model needs to choose from four candidate complements that can be consistent with question in commonsense. Part of the training data and all validation data comes from original CODAH. They also include 7k SWAG validation examples as additional training data.

4.2 Baselines

For X-CODAH and X-CSQA datasets, we mainly compare with the previous state-of-the-art MCP (XLMR-L) (Lin et al., 2021) as well as other three multilingual pretrained langauge models: mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) base and large models. For MCP (XLMR-L), they first create a multilingual parallel dataset MickeyCorpus from OMCS which has 561k sentences in 11 languages. Then based on XLM-RoBERTa large model, they first fine-tune on the reformated multiple-choice question answering dataset MickeyCorpus (Lin et al., 2021) and further fine-tune on the final datasets X-CODAH and X-CSQA.

4.3 Implementation Details

We use Microsoft Machine Translator [3] for all translations, including translating the given query, the retrieved knowledge and English training data to other 15 languages. We will release these translations for academic usage. For Wiktionary, we use the dump of Wiktionary which includes 999,614 definitions. We empirically obtaining 6 words definitions from Wiktionary for X-CODAH (see Figure 3 (a)) and use the provided question concept and answer as two candidate words for X-CSQA. For ConceptNet, we use ConceptNet version 5.7.0 [4]. For GPT-3, we use the curie [5] model.

Our model implementation is based on HuggingFace's Transformers Library (Wolf et al., 2020). We conduct all experiments on 8 Nvidia V100-32GB GPU cards. We follow the configurations in XCSR to pretrain the MCP model based on XLM RoBERTa large except that the maximum sequence length is 256 and batch size is 32. The accuracy of the resulting MCP checkpoint on its dev set is 87.4. We then initialize with this checkpoint for further fine-tuning with the extracted knowledge from different knowledge sources. During fine-tuning, we set the training epochs, batch size and gradient accumulation steps as 10, 4 and 2 respectively. The total batch size here is 64 by "*batch size per device* × *# GPUs* × *# gradient accumulation steps*". For hyper-parameter search, we sweep over the learning rates $\in \{1e-5, 3e-5, 5e-5, 3e-6, 5e-6\}$ and report the maximum results.

4.4 Experimental Results

Results on test set Table 4 summarizes our results on the hidden test set from XCSR leaderboard. TRT outperforms all previous works by a significant margin on both datasets, achieving the average score of 59.8/63.7 with an absolute improvement of 3.3/3.6 over previous state-of-the-art MCP(XLMR-L). For some high-resource languages, like Ger-

[3]https://azure.microsoft.com/en-us/services/cognitive-services/translator/

[4]https://github.com/commonsense/conceptnet5

[5]https://beta.openai.com/pricing

Dataset	Model	en	de	it	es	fr	nl	ru	vi	zh	hi	pl	ar	ja	pt	sw	ur	avg
Zero-shot transfer (models are trained on English data) and evaluate on the target language																		
X-CODAH	MCP (XLMR-L)	69.7	63.0	62.3	63.0	64.7	64.7	55.0	55.0	59.7	54.3	61.7	52.3	57.0	55.0	40.3	49.3	57.9
	+ Wikt.	72.0	65.3	63.0	65.0	66.0	66.0	58.7	59.3	58.0	54.3	64.0	55.7	61.3	60.7	47.0	53.0	**60.6**
	+ Cpnt.	72.3	68.3	65.7	65.0	66.0	64.3	60.3	57.0	58.3	55.0	65.3	53.7	57.3	59.7	46.3	52.0	60.4
	+ OMCS	73.0	67.0	64.0	63.7	63.0	62.0	57.3	60.0	62.0	53.0	63.7	56.0	57.7	59.3	44.0	49.3	59.7
	+ GPT-3	71.7	62.0	64.3	62.3	65.0	62.3	56.7	55.3	58.0	54.3	64.7	55.0	59.3	60.0	42.7	52.7	59.1
X-CSQA	MCP (XLMR-L)	69.0	57.6	57.2	57.9	59.9	56.1	55.2	56.0	56.6	48.8	56.4	52.5	50.8	58.3	42.5	47.4	55.1
	+ Wikt.	70.7	59.5	60.2	61.4	59.5	58.5	56.6	55.6	58.3	51.2	56.0	55.6	52.0	60.6	46.8	49.1	**57.0**
	+ Cpnt.	70.7	57.2	58.1	58.6	58.7	55.8	55.5	56.0	56.6	49.9	55.9	53.9	52.4	55.6	43.3	47.8	55.4
	+ OMCS	70.5	59.9	59.3	60.5	60.0	56.8	55.3	56.1	57.3	48.9	56.4	53.4	51.6	59.0	46.7	48.0	56.2
	+ GPT-3	70.3	57.2	58.8	60.2	58.3	58.1	54.8	55.0	55.6	49.0	54.5	52.9	52.1	57.9	42.9	47.6	55.3
Translate-train (models are trained on English training data and its translated data) and evaluate on the target language																		
X-CODAH	MCP (XLMR-L)	71.0	70.7	66.3	69.7	70.7	66.7	63.7	62.3	62.3	60.3	64.7	59.3	59.7	67.7	57.0	57.7	64.4
	+ Wikt.	72.0	71.7	68.0	69.3	69.7	67.0	65.3	66.0	63.0	61.0	65.0	58.3	62.7	68.0	58.0	58.3	65.2
	+ Cpnt.	70.7	68.7	67.0	68.0	68.0	68.3	65.0	62.0	61.7	56.3	65.0	61.7	62.3	66.3	60.0	57.3	64.3
	+ OMCS.	74.7	69.7	67.3	67.7	67.7	68.3	62.7	65.3	65.3	58.7	68.3	62.0	64.0	68.3	56.7	59.7	**65.4**
X-CSQA	MCP (XLMR-L)	69.4	59.3	60.6	60.9	60.8	57.9	57.0	58.2	58.0	50.4	58.3	55.1	53.9	60.3	47.1	50.9	57.4
	+ Wikt.	70.0	61.7	61.2	61.1	60.9	59.8	59.8	59.3	59.6	53.8	59.7	58.1	54.3	60.5	51.8	52.8	**59.0**
	+ Cpnt.	68.5	59.2	59.5	58.2	61.3	58.7	56.6	57.9	58.3	52.6	58.4	55.6	52.9	60.5	48.2	52.8	57.4
	+ OMCS	71.7	61.1	63.6	62.8	60.3	58.6	58.1	59.3	58.5	51.7	58.1	56.1	54.2	60.4	48.6	53.4	58.5

Table 5: Comparisons for TRT with different knowledge sources in the zero-shot transfer and translate-train setting on the development set. Wikt. and Cpnt. are short for Wiktionary and ConceptNet. Results of GPT-3 are by using the generated knowledge with prompt from Wikitionary.

MCP (XLMR-L)	+ G-Wikt	G-Cpnt.	G-OMCS
57.9	59.1	58.2	58.4

Table 6: Zero shot results with GPT-3 in different knowledge formats on X-CODAH. G-Wikt., G-Cpnt. and G-OMCS. are short for using GPT-3 to generate definition, triple and sentence as the knowledge format in Wiktionary, ConceptNet and OMCS.

man (de), we observe larger gain with 4.6 points improvement on X-CSQA. For low-resource languages, like Swedish, there are even larger gains with 7.7 and 7.9 improvements on X-CSQA and X-CODAH.

Effectiveness of different knowledge sources Table 5 list the detailed comparisons among different knowledge sources in both zero-shot and translate-train settings. We observe the following findings from these results: (i) Knowledge can be helpful for multilingual commonsense reasoning in both zero-shot and translate-train setting. For example, in the zero-shot setting, TRT with Wiktionary improve 2.7 and 1.9 points over the MCP (XLMR-L) baseline on X-CODAH and X-CSQA. In translate-train setting, there are 1.0 and 1.6 improvements for each dataset. (ii) Wiktionary helps the most among all knowledge sources in both settings, except that OMCS performs slightly better than Wiktionary on X-CODAH in the translate setting. We hypothesize that the difficulty of understanding hardness words can be mitigated by incorporating additional knowledge as context. (iii) The generated knowledge from GPT-3 can also improve over the baseline, without leveraging machine translation and explicit knowledge, which demonstrates the rich implicit knowledge in GPT-3. For example, for X-CODAH dataset, GPT-3 can outperform the baseline about 1.2 point. However, there still exist the gap between GPT-3 and designated knowledge format. We leave this one as future work to bridge the gap.

Effectiveness of GPT-3 generation in different knowledge formats As GPT-3 requires the prompt with input and output examples, we feed it with a few examples with a query and the knowledge in three designated formats from Wiktionary, ConceptNet and OMCS on X-CODAH. Table 6 shows the zero-shot results with different generated knowledge from GPT-3. We observe that G-Wikt. outperforms the baseline MCP (XLMR-L) 1.2 points while G-Cpnt. and G-OMCS don't show significant improvements. This demonstrates GPT-3 does exist implicit knowledge in its parameters. But the generated knowledge from GPT-3 is still less helpful than the same knowledge format from the explicit Wiktionary which indicates the potential improvement with the language model. To further look at the generated knowledge, Table 7 lists five examples for G-Wikt., G-Cpnt, and G-OMCS respectively. We can see that most of them make sense and especially the quality of generated knowledge from G-Wikt. looks good. These also

Question	G-Wikt.	G-Cpnt.	G-OMCS
Swat officers sweep the space with rifle lights. Someone climbs backward through the narrow vent hole.	sweep: To clean by means of a stroking motion of a broom or brush.	sweep has context card games.	vent-hole is a synonym of vent.
A boy is running across a field wearing a green shirt. He smiles because his shirt is bright green.	shirt: A piece of clothing worn by men and women.	field defined as same shape as ribbon	The boy is wearing a green shirt.
The dog stands to catch the Frisbee the leans on the man. The dog jumps into the man's arms.	lean: To rest on something.	dog defined as animal	The dog jumps into the man's arms.
We see a colorful and playful title screen. We then see people in a room and outdoors at a fancy party.	fancy: Showy or pretentious.	title screen defined as same shape as ribbon	The title screen is colorful and playful.
Someone glares at the stick then at someone. Someone leans the stick against the bed.	glare: To direct a look of anger or hatred at someone.	glare similar to look	The stick is leaning against the bed.

Table 7: GPT-3 generation examples in different knowledge format. G-Wikt., G-Cpnt. and G-OMCS are short for the GPT-3 generated knowledge with prompts from Wiktionary, ConceptNet and OMCS.

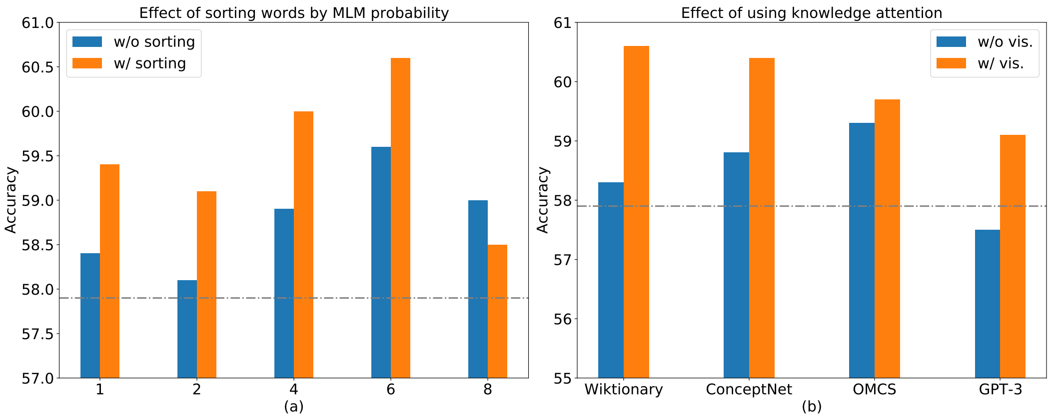

Figure 3: Effects of the number of word definitions and the visible knowledge attention mechanism on X-CODAH dataset. Figure (a) shows the performance can be improved by selecting the hardness words and increasing the number of definitions from 1 to 6. Figure (b) shows visible knowledge attention can be helpful on a variety of knowledge sources. The dashed lines in figure (a) and (b) represent the baseline result.

explain the larger improvement with G-Wikt. than G-Cpnt. and G-OMCS.

Effectiveness of sorting definitions by MLM probability In Section 3.3, we introduce using masked language model (MLM) to select the top-N hardness words with the lowest probability. Here we perform an ablation study by comparing this strategy (w/ sorting) with randomly choosing the words (w/o sorting). As shown in Figure 3 (a), sorting by MLM probability can outperform the random selecting, especially with a smaller number of words, achieving the best performance with 6 words definitions. However, there is no much difference when we use eight definitions.

Effectiveness of knowledge attention In Section 3.4, we mention that simply appending knowledge as additional context can be noise to some tasks like X-CODAH, a scene completion tasks, which may divert the original semantic meaning. Therefore, here we compare the model performance between full attention (w/o vis.) and visible knowledge attention (w/ vis.) on all investigated knowledge sources (Wiktionary, ConceptNet, OMCS and GPT-3). As shown in Figure 3 (b), visible knowledge attention can consistently outperform full attention on all knowledge sources. For example, there are 2.3 and 1.6 points improvement between them when integrating from Wiktionary and GPT-3.

5 Conclusion

In this work, we first present the translate-retrieve-translate (TRT) strategy for multilingual commonsense reasoning that collects related knowledge via translation and then retrieval from the knowledge sources. Then we conduct extensive experiments by utilizing a diverse of four English knowledge sources, including Wiktionary, ConceptNet, OMCS and GPT-3. By using TRT with different knowledge sources, we achieve state-of-the-art results on XCSR leaderboard which demonstrates the effectiveness of our proposed methods. Future work includes more effective ways to incorporate the diverse knowledge sources into pre-training and fine-tuning stage for commonsense reasoning.

References

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating external knowledge into machine reading for generative question answering. 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 2521–2530, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1535–1545.

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)*, pages 1295–1309, Online. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2021. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. *arXiv preprint arXiv:2109.12573*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021a. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *to appear in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021b. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

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)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. 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 1274–1287, Online. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. To appear.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*, pages 2901–2908.

Peter LoBue and Alexander Yates. 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*, pages 329–334.

Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From freebase to wikidata: The great migration. In *Proceedings of the 25th international conference on world wide web*, pages 1419–1428.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

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

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

Push Singh. 2002. The open mind common sense project. *KurzweilAI. net*.

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*.

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, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Colin Wei, Sang Michael Xie, and Tengyu Ma. 2021. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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.

Ruochen Xu, Yuwei Fang, Chenguang Zhu, and Michael Zeng. 2021a. Does knowledge help general nlu? an empirical study. *arXiv preprint arXiv:2109.00563*.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021b. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207, Online. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*.