Commonsense Knowledge Base Population and Reasoning for Inferential Knowledge

by

Tianqing Fang

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in Computer Science and Engineering

August 2024, Hong Kong

AUTHORIZATION

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Tianqing Fang

23 August 2024

Commonsense Knowledge Base Population and Reasoning for Inferential Knowledge

by

Tianqing Fang

This is to certify that I have examined the above PhD thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Yangqiu Song, Thesis Supervisor

Prof. Xiaofang Zhou, Head of Department

Computer Science and Engineering
23 August 2024

ACKNOWLEDGMENTS

I express my deep gratitude to my supervisor, Prof. Yangqiu Song, for his invaluable support throughout my Ph.D. journey. I also extend my thanks to Prof. Antoine Bosselut from EPFL and Prof. Muhao Chen from UC Davis for their mentorship during my visit and profound insights into Natural Language Processing and Artificial Intelligence.

I am grateful to my thesis proposal committee and thesis defense committee, Prof. Junxian He, Prof. Brian Mak, Prof. Dan Xu, Prof. Jing Wang, Prof. Wen Hua, and Prof. Irene Lo, for their participation and valuable feedback on my thesis.

I would like to thank all of my co-authors, Dr. Hongming Zhang, Weiqi Wang, Dr. Bin He, Prof. Nedjma Ousidhoum, Xinran Zhao, Prof. Dit-Yan Yeung, Haojie Pan, Dr. Xin Liu, Jiefu Ou, Haowen Ke, Dr. Kun Xu, Dr. Dong Yu, Sehyun Choi, Shibo Hao, Prof. Ziqian Zeng, Weimin Ni, Xiang Li, Mutian He, Zhaowei Wang, Dr. Ginny Y. Wong and Simon See, Ying Su, Zihao Wang, Prof. Tong Zhang, Quyet V. Do, Jiayao Zhang, Baixuan Xu, Chun Yi Louis Bo, Prof. Lei Chen, Haochen Shi, Wenxuan Ding, Dr. Shizhe Diao, Yongyu Lei, Dr. Liangming Pan, Wangchunshu Zhou, Sedrick Scott Keh, Prof. Min-Yen Kan, Jiayang Cheng, Lin Qiu, Tsz Ho CHAN, Chunkit Chan, Jiaxin Bai, Qipeng Guo, Prof. Yue Zhang, Prof. Zheng Zhang, Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Mingxi Cheng, Prof. Paul Bogdan, Yuxin Jiang, Dr. Wenxuan Zhou, Dr. Fangyu Liu, and Zeming Chen. Thank you all for your support in enabling our publications. The order of names is based on the publication dates of our co-authored papers.

Among all my co-authors, I especially want to highlight Dr. Hongming Zhang, whose support has been invaluable throughout my Ph.D. journey and most of my research projects. I also greatly appreciate his recommendations during my job search. I would also like to acknowledge Mr. Weiqi Wang, a junior Ph.D. student whom I have mentored since his undergraduate years. Thank you for your hard work on conceptualization and all the others.

There are other (Ph.D) researchers where I have worked together but may not have published together, such as Huiru Xiao, Yan Liang, (also Ziqian Zeng and Xin Liu which have been mentioned as co-authors) with whom we had wonderful time during the pandemic in Hong Kong, and Qin Liu, Fei Wang, Bangzheng Li, Tenghao Huang, James Huang, from USC, and Silin Gao, Saibo Geng, Deniz Bayazit, Beatriz Borges, Dr. Syrielle Montariol, Sepideh Mamooler, Badr AlKhamissi, Dr. Debjit Paul, Mete Ismayilzada, Anna Dai, from EPFL, and Dr. Kaixin Ma, Dr. Zhisong Zhang, Dr. Wenhao Yu, Peng Zhou from Tencent. I also thank other group members in HKUST, Zizheng Lin, Yue Guo, Haoran Li, Qi Hu, and Wei Fan.

I am grateful to my old friend Yuyan Li for the constant support and discussions over the years. I also thank my friends Banruo Xianyu, Wenxuan Hu, and Yuxuan Chen for their spiritual support. My heartfelt appreciation goes to Luzhe Huang and Xiongye Xiao, friends since undergrad, for the wonderful times in LA. I also thank Xiao Pan and Jianyang Gu for our frequent discussions about the academic futures. Finally, I want to express my gratitude to my roommates Yiming Fu, Shihan Zhu, and Zhiyuan Yu. It was comforting to have you with me during the challenging times of the pandemic.

I also thank the support from my girlfriend Xiaohui Jiang, and my parents. Thank you for giving me the courage to pursue my academia dream!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Commonsense Knowledge Base Population and Reasoning for Inferential Knowledge

by

Tianqing Fang

Computer Science and Engineering

The Hong Kong University of Science and Technology

ABSTRACT

Commonsense knowledge includes facts about the everyday world that ordinary people are expected to know. It plays a crucial role in natural language processing (NLP) systems, enabling them to make presumptions about common situations encountered by humans. However, acquiring and incorporating commonsense knowledge into NLP systems poses challenges, as such knowledge is typically implicit and not readily available in standard corpora.

To tackle the data scarcity issue, a standard way to study commonsense is to construct commonsense knowledge bases (CSKBs). Previous attempts have focused on (1) human annotation, which is expensive and has limited scalability; (2) information extraction, which suffers from relatively poor quality and reporting bias; or (3) text generation from Large Language Models (LLMs), which suffers from selection bias and limited novelty of generated knowledge. Moreover, the power of LLMs to elicit commonsense knowledge also requires fine-tuning on large-scale corpora and human-annotated commonsense data in the first place.

We propose an alternative commonsense knowledge acquisition framework, called Commonsense Knowledge Base Population (CKBP), which automatically populates complex commonsense knowledge from more affordable linguistic knowledge resources. We establish a benchmark for CKBP based on event-event discourse relations extracted through semantic and discourse parsing of large corpora, and we manually annotate 60K populated triples for verification.

To carry out the population process, we introduce a Graph Neural Network (GNN)-based model that leverages the rich contextual information in the knowledge graph as additional supervision signals. Since CKBP is a semi-supervised learning problem with a large amount of unlabeled data (discourse knowledge from large corpora), we also propose a pseudo-labeling-based model that achieves excellent performance. We evaluate the effectiveness of the populated knowledge on downstream commonsense reasoning tasks and observe that it enhances generative commonsense inference and commonsense question answering by providing more diverse knowledge.

Furthermore, with the knowledge at hand, we explore commonsense reasoning based on commonsense knowledge from two perspectives. First, we directly utilize the populated knowledge for downstream commonsense question answering by converting it into question-answering (QA) form with templates, serving as supervision data for training QA models and generative commonsense inference models. Second, we perform reasoning on complex logical queries derived from commonsense knowledge graphs. We sample conjunctive logical queries from the knowledge graphs and verbalize them using LLMs to generate narratives for both training and evaluating models for complex reasoning. Experimental results demonstrate that while LLMs exhibit proficiency in handling one-hop commonsense knowledge, performing complex reasoning involving multiple hops and intersections on commonsense knowledge graphs remains challenging. Models trained on complex logical queries show improvement in terms of general narrative understanding and complex commonsense reasoning ability.

CHAPTER 1

INTRODUCTION

1.1 Commonsense Knowledge

1.1.1 Definition of Commonsense Knowledge

Commonsense knowledge is a collection of information about everyday things that all humans are expected to know [12], like knowledge regarding entities such as “lemons are sour”, and event-event inferential knowledge such as “if someone gets hungry and then he/she wants to eat”. It is also defined or characterized as “commonsense knowledge includes the basic facts about events (including actions) and their effects, facts about knowledge and how it is obtained, facts about beliefs and desires” by John McCarthy in 1989 [13], and “what a typical seven year old knows about the world” by Ernest Davis [14].

Commonsense knowledge is crucial for natural language processing systems to understand human language. For example, a recommendation system needs to know that *lemon is sour* in order to recommend food to customers to fit their interests. In dialogue systems, social commonsense knowledge and reasoning are also needed for machines to understand human interactions better [15, 16]. When a driver tells his speech assistant that he is *out of gas*, the system should be equipped with the ability to infer that the driver would like to find a gas station nearby. With the rapid development of artificial intelligence systems, a large amount novel datasets and benchmarks are created, leading to an emerging field of commonsense reasoning, including but not restricted to domains such as daily entity understanding [17, 18, 19], social interaction [20, 21], spatial and temporal events [22, 23, 24], numerical understanding [25, 26], stories and narratives understanding [27, 1], and visual commonsense understanding [28, 29, 30].

Two pioneer works of computational commonsense knowledge are Cyc [31] and ConceptNet [12]. Researchers propose to canonicalize world knowledge into knowledge bases in order to collect commonsense knowledge in the open world that machines do not possess. Cyc is a commonsense knowledge base spanning everyday objects and actions like *You have to be awake to eat*. While Cyc requires commercial licenses, ConceptNet is an open-source commonsense knowledge base defined across curated relations. With several rounds of development, ConceptNet 5.7 [32] now contains 3.4 Million entity-centric tuples in English by crowdsourcing and merging with existing large-scale knowledge bases, such as DBPedia [33] and WordNet [34]. With the rapid development of NLP techniques, more is needed to investigate only object-level commonsense defined in ConceptNet. ATOMIC is then collected across nine relations investigating social causes and effects, consisting of 880K tuples, e.g., ("If X repels Y's attack", `Effect on Y`, "Y arrested by the police"), representing day-to-day inferential knowledge. Such everyday if-then commonsense relations are crucial for machines to understand human language.

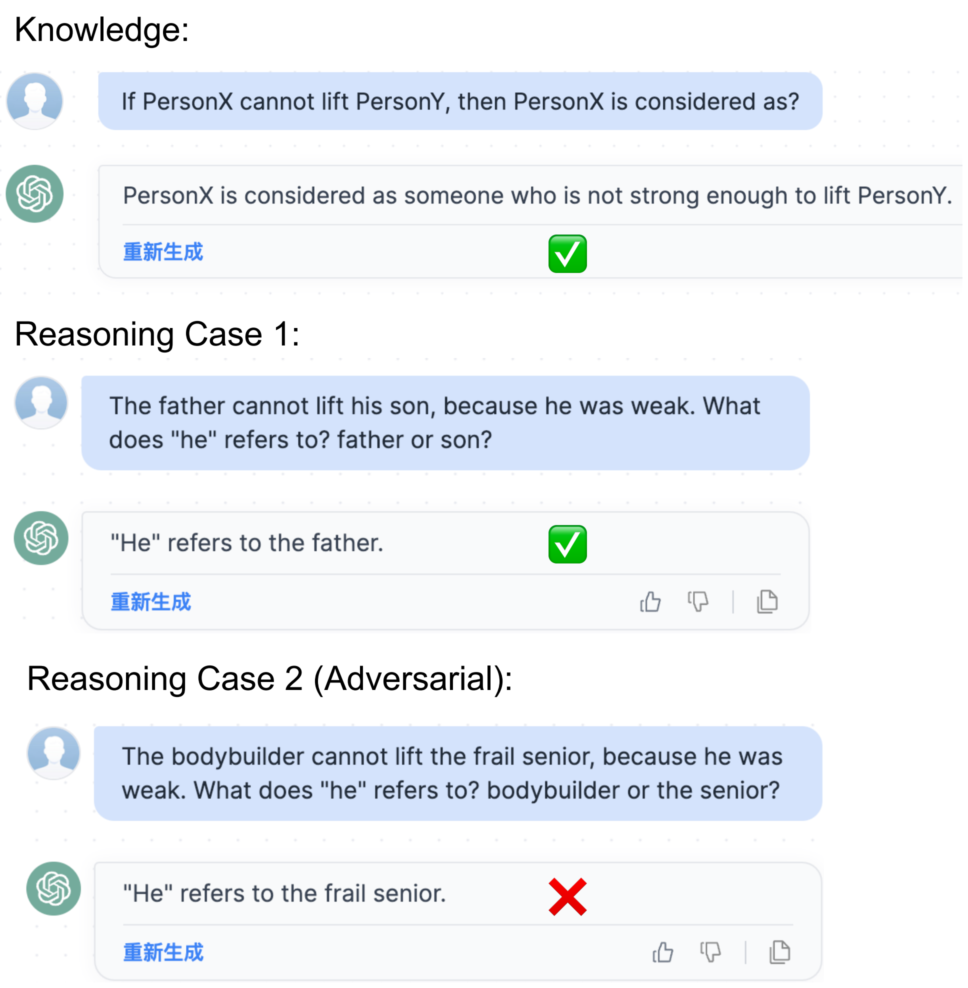
In this thesis, we mainly focus on event-level inferential knowledge (defined in ATOMIC), which has three advantages. First, the ATOMIC paradigm focuses on events, situations, and states, which represent more complicated components than only entities and can have a broader range of applications. Second, the nodes in ATOMIC are represented as free text, and relations can be easily converted to natural language. This loosely structured form aligns well with the current go-to backbone for performing natural language processing, the Large Language Models (LLMs). Third, the ATOMIC paradigm relies on humans to annotate plausible commonsense, which is subjective and intuitive. Unlike formal logic, it better aligns with the unique characteristic of commonsense knowledge.

1.1.2 Limitations and Challenges for Commonsense Acquisition and Reasoning

It is previously assumed that language models can only capture limited commonsense knowledge by only self-supervised learning (models such as BERT [35], GPT [36], and GPT2 [37], and T5 [38]), as commonsense knowledge is usually not explicitly presented in the training corpora. The capacity of those Pre-trained Language Models (PLMs) is not large enough to discover complex commonsense knowledge [39, 40]. This drives the need to curate large-scale commonsense knowledge resources for “small” PLMs to equip them with the ability to use commonsense reasoning.

With the rapid development of backbone language models, modern Large Language Models (such as ChatGPT [41], Llama [42], Mistral [43], and Qwen [44]) have shown a remarkable performance on many reasoning benchmarks [45, 46, 47, 48], yet there still exists a need to ensure the alignment between the generation of LLMs with commonsense knowledge to avoid hallucination and for safer usage against social bias [49, 50, 51]. This motivates us to curate high-quality, large-scale, and high-coverage commonsense knowledge resources that can be used to perform more accurate reasoning. However, existing commonsense resources are either of a small scale and coverage (human annotation), of a large scale but poor quality (information extraction), or of a large scale but not novel (language model generation), due to the high cost of human annotation [12], the quality issue and *reporting bias* [52] of information extraction, and *selection bias* [53] in language models.

In addition, despite those LLMs having a large capacity to store numerous knowledge, they may still fall short in terms of reasoning based on knowledge [54]. For instance, [54] reported that ChatGPT can generate reasonably relevant knowledge when provided with appropriate prompts for a commonsense reasoning task. However, despite its existing knowledge, ChatGPT may not always provide correct answers when faced with concrete reasoning scenarios in a different context. Similar situations are also observed for multi-hop or complex commonsense reasoning, where LLMs fall short in performing conjunction, projection, and negations over existing “known” knowledge (Chapter 7). An example derived from the famous Winograd Schema Challenge is provided in Figure 1.1. The original reasoning question is Reasoning Case 1, where a coreference resolution process is required to determine what “he” refers to given the context: “ The father could not lift the son because he was weak.” Here, the knowledge needed is that if someone X cannot lift someone Y, then X is considered weak. Even though ChatGPT already “knows” such knowledge, as shown in the first sub-figure, it makes a mistake if we replace “father” with a “bodybuilder” and “son” with a “frail senior”. This is because the “frail senior” is more semantically asso-



Knowledge:

If PersonX cannot lift PersonY, then PersonX is considered as?

PersonX is considered as someone who is not strong enough to lift PersonY. ✓

Reasoning Case 1:

The father cannot lift his son, because he was weak. What does "he" refers to? father or son?

"He" refers to the father. ✓

Reasoning Case 2 (Adversarial):

The bodybuilder cannot lift the frail senior, because he was weak. What does "he" refers to? bodybuilder or the senior?

"He" refers to the frail senior. ✗

Figure 1.1: A case when ChatGPT possesses the correct knowledge but fails to perform the correct reasoning.

ciated with “weak”, thus being more confusing for statistical language models to perform reasoning.

To tackle the above limitations, we propose Commonsense Knowledge Base Population (CKBP), a framework that can scalably harvest commonsense knowledge of high quality and high novelty without expensive human annotation, to deal with the limitation of not having enough commonsense knowledge. Second, to tackle the reasoning issue of commonsense knowledge, we propose a framework to sample large-scale complex reasoning signals from commonsense knowledge bases based on conjunctive logical queries.

1.2 Commonsense Knowledge Population and Reasoning

In this thesis, we first propose the task formulation and benchmarking of a new commonsense acquisition paradigm, Commonsense Knowledge Base Population. Then, we introduce novel algorithms for performing Commonsense Knowledge Base Population. Third, we introduce how we leverage the (populated) commonsense knowledge to improve the problem solving and reasoning ability of (large) language models.

1.2.1 Benchmarking Commonsense Knowledge Base Population

First, we introduce our Commonsense Knowledge Base Population problem formulation and benchmarking.

Throughout the development of automated commonsense understanding, CommonSense Knowledge Base (CSKB) is an important form of automatic commonsense reasoning system to store knowledge sources for drawing inferences. With expert-curated relations and human annotations, CSKBs such as ConceptNet [12], ATOMIC [55, 10], and GLUCOSE [1] are developed to study commonsense regarding properties of objects, causes and effects of events and activities, motivations and emotional trajectories of humans on specific circumstances, and so on.

First, the knowledge acquired by those CSKBs is based on crowdsourcing, which is relatively more expensive than other automatic information extraction methods. To scalably acquire new knowledge, COMET [56] is proposed to finetune a large pre-trained language model (i.e., GPT [36]) with existing commonsense knowledge bases (for example, ATOMIC) such that they can automatically generate reasonable commonsense knowledge. Even though COMET can generate high-quality commonsense knowledge with the supervised learning approach, it tends to fit the training data too well to generate novel concepts and knowledge. This is usually called a *selection bias* problem in statistical analysis [53, 57].

On the other hand, although information extraction may also be subject to *reporting bias* [52], where the frequencies may not truly reflect the relative likelihood in the real world, it can provide many candidate examples that can be evaluated by a machine learning model trained on human-annotated data. Such an automatic extraction approach can easily scale up to two orders of magnitude larger than human annotations.

To tackle those limitations, we propose to use information-extraction tools to extract coarse sentence-level discourse relations first. These discourse relations, such as extracting ("I have lunch", `Precedence`, "I am full") from a sentence "I have lunch and then I am full", can serve as natural sources to derive inferential commonsense knowledge such as "If someone X has lunch, then X will be full". We use an off-the-shelf knowledge graph, ASER [58, 59], that encompasses billion-scale discourse knowledge, as the source discourse knowledge, and unify the format of ASER with the format of the target commonsense knowledge base, ATOMIC. Then, the Commonsense Knowledge Base Population (CKBP) framework adopts a classifier that is finetuned with gold-standard human-annotated commonsense knowledge to discriminate whether a discourse knowledge can be converted to a corresponding commonsense knowledge. The populated results are evaluated based on accuracy (quality), novelty (how much new knowledge can be acquired), and diversity.

1.2.2 Reasoning for Commonsense Knowledge Base Population

We present several methods of modeling this classifier, including graph-aware language modeling, graph neural networks, and semi-supervised learning.

Although it is widely accepted that the graph substructure can be helpful in making predictions and inferences in entity-centric knowledge graphs [60], existing commonsense knowledge-based models [56, 40] still treat the prediction as a translational problem for the triplets in the knowledge base and do not consider the subgraph structures. It is also not trivial to leverage graph structures in commonsense knowledge acquisition. First, there is no existing graph structure in the CSKBs, as the labeling procedure only considers the head, the tail, and their relations. For example, the average degree in ConceptNet and ATOMIC is ten times smaller than that of regular factual knowledge bases such as Freebase [2]. There are few overlaps between heads and tails, given that both can be arbitrary texts. The heads and tails can form a bipartite graph, but graph convolution in such a graph may not provide additional information compared to direct representation learning for nodes because tails can be conditionally independent given a head. However, with ASER, a more structural knowledge graph, it is possible to perform more complicated reasoning over the substructures. Second, as we mentioned, heads and tails are loosely structured texts in both ATOMIC and ASER, and a contextualized representation model should be applied to them for better representations. We developed two models, BERTSAGE and KG-BERTSAGE, which leverage the semantic representation of nodes or edges and graph structures to aggregate contextual information.

Besides graph structure, another dominant feature of the CSKB Population is the imbalance between labeled data (around one million) and unlabeled data (over two hundred million). Moreover, as CSKBs inherently provide only ground-truth (positive) examples, the randomly sampled negative examples in the task are less informative and may lead the model to overfit artifacts of the dataset. A supervised learning model finetuned on such an annotated training set is hard to generalize to out-of-domain knowledge space. Therefore, it is natural to develop a semi-supervised learning algorithm to provide pseudo-positive and negative labels. We introduce PsuedoReasoner, a pseudo-label-based semi-supervised learning algorithm that involves a unique data filtering process and can significantly improve population performance.

1.2.3 Using Commonsense Knowledge for Downstream Reasoning

After acquiring more commonsense knowledge, there are several ways to improve downstream commonsense reasoning, such as verbalizing the knowledge into question-answering pairs for training a question-answering model and regarding the knowledge as input text for auto-regressive finetuning. We tested our populated knowledge in both settings and found that the knowledge acquired by CKBP can significantly improve commonsense question answering and generative commonsense inference compared to only using human-annotated commonsense knowledge.

Large language models need help to effectively perform reasoning when presented with complex tasks, such as reasoning about multiple events and their relationships. This shortcoming is due to the inherent difficulty of reasoning over multiple pieces of information and a lack of adequate-scale, supervised training datasets for learning [61]. Unfortunately, complex and multi-hop commonsense reasoning benchmarks [9] are both technically challenging and financially expensive to curate. Consequently, previous efforts either constructed datasets (a) with simpler reasoning structures, such as single-hop chains [62], (b) using distant supervision based on one-hop inference [9], or (c) with human-annotations, but at a relatively small scale [8].

In this thesis, we construct COM2 (**COM**plex **COM**monsense), a novel commonsense reasoning dataset using multi-hop queries in commonsense knowledge graphs to construct question answer pairs requiring complex narrative reasoning. To build this dataset, we use *conjunctive logical queries* [63], a subset of First-Order Logical queries that use existential quantifiers and conjunction. The multi-hop projection operation involves inferring hidden contexts, while the intersection operation enables reasoning among multiple events, encompassing common cause or effect and abduction. We apply those complex reasoning data to instruction tuning for large language models to enhance models' complex reasoning ability significantly.

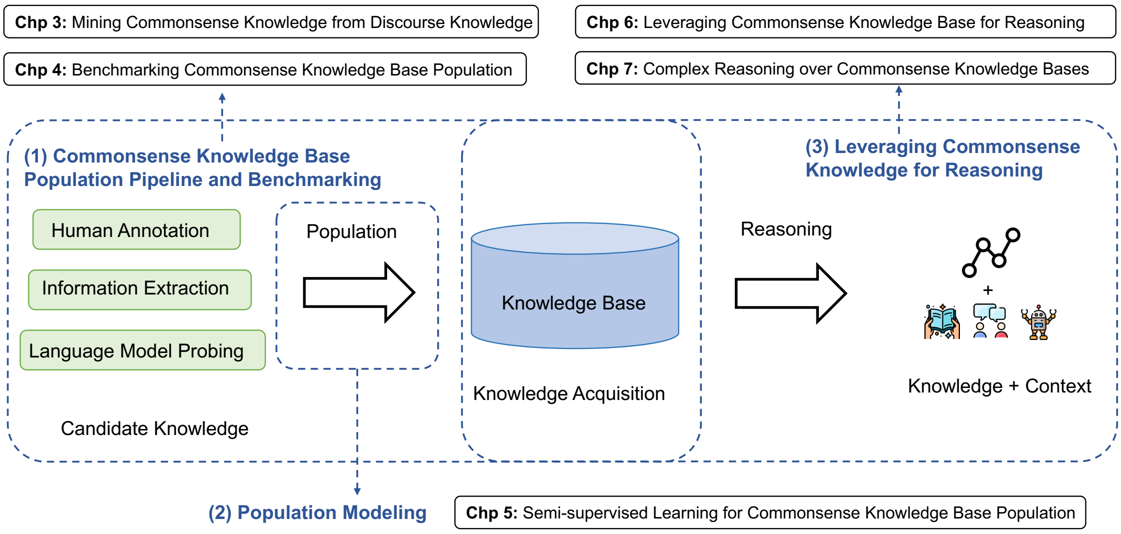

Chp 3: Mining Commonsense Knowledge from Discourse Knowledge
Chp 4: Benchmarking Commonsense Knowledge Base Population
Chp 6: Leveraging Commonsense Knowledge Base for Reasoning
Chp 7: Complex Reasoning over Commonsense Knowledge Bases
(1) Commonsense Knowledge Base Population Pipeline and Benchmarking
Human Annotation
Information Extraction
Language Model Probing
Candidate Knowledge
Population
Knowledge Base
Knowledge Acquisition
Reasoning
(3) Leveraging Commonsense Knowledge for Reasoning
+
Knowledge + Context
(2) Population Modeling
Chp 5: Semi-supervised Learning for Commonsense Knowledge Base Population

Figure 1.2: Roadmap of the thesis.

1.3 Thesis Organization

As shown in Figure 1.2, this thesis is organized as follows. In Chapter 2, we introduce the relevant works regarding commonsense acquisition, commonsense reasoning, and complex reasoning. In Chapter 3, we present the first work of Commonsense Knowledge Base Population, which builds fundamentals of converting discourse knowledge to commonsense knowledge. Such a pipeline can provide accurate, novel, and diverse commonsense knowledge at scale without extra human annotation efforts. In Chapter 4, we benchmark the process of the Commonsense Knowledge Base Population by unifying four popular commonsense knowledge bases of different topics to make it comprehensive. We provide a human-annotated evaluation set of around 30K examples to evaluate the models' commonsense population ability. On top of this, a Graph Neural Network-enhanced method is also introduced to improve structured commonsense modeling. In Chapter 5, we investigate CKBP from the angle of semi-supervised learning, as the already human-annotated CSKB can be considered as labeled data, and that information-extracted candidate knowledge is natural unlabeled data. We build a pseudo-label-based semi-supervised learning framework with a quality filter and influence function filter to improve the quality of pseudo labels. Chapter 6 introduces the experiments on leveraging populated knowledge for downstream commonsense question answering and generative commonsense inference. Next, in Chapter 7, we go beyond the Commonsense Knowledge Base Population and investigate reasoning based on Commonsense Knowledge Bases. We formally define complex reasoning based on logical queries on CSKBs and study how language models perform on this complex reasoning task.

CHAPTER 2

RELATED WORKS

This thesis focuses on how to effectively acquire commonsense knowledge at scale and leverage it to enhance the reasoning ability of backbone language models. We thus introduce related works regarding commonsense acquisition (Section 2.1), a non-trivial task because commonsense is usually implicit and often omitted in human language for efficient communication [64]. Then, we introduce machine commonsense reasoning (Section 2.2), including several typical task formats such as commonsense knowledge base reasoning, commonsense question answering, and leveraging commonsense to help real-world tasks. Besides one-hop commonsense reasoning, we also study complex reasoning (Section 2.3), which includes multiple rounds or hops in the reasoning chain.

2.1 Commonsense Acquisition

AI relies heavily on commonsense knowledge, which is essential for its functioning. The initial step in studying commonsense knowledge is acquiring it. There are three main approaches to acquiring commonsense knowledge, which we will outline below, namely human annotation, information extraction, and language model generation. This thesis focuses on leveraging the advantages of all methods. This includes using information extraction to acquire large-scale candidate commonsense knowledge, using human-annotated commonsense knowledge as seeds to train a language-model-powered commonsense classifier, and harvesting commonsense knowledge at scale.

2.1.1 Human Annotation

Textual Commonsense Human annotation has been a critical part of collecting commonsense, especially Commonsense Knowledge Bases (CSKB), starting from early pioneer works, such as logical formulae in Cyc [31, 65] or textual assertions in OMCS [66], which is later utilized as a core part of the renowned ConceptNet [12, 32] focused on entity-centric knowledge. There are also attempts to annotate event-centric/situational knowledge, with focuses on social commonsense [67, 68]), dialogue [69, 70], and narratives [62], as well as ATOMIC [55] and ATOMIC-2020 [10] with a range of dimensions of if-then reasoning on events.

In addition to these flat and one-hop commonsense knowledge formulated in triples, there is another kind of commonsense representation using conceptualization or abstraction [71, 72]. Conceptualization basically performs a contextualized IsA relation mapping to derive higher-order knowledge, for example, deriving "stimulants will refresh people" from "coffee will refresh people" by conceptualizing "coffee" to a kind of "stimulant." The candidates of conceptualized knowledge are usually derived by linking entities or events to conceptual knowledge bases such as WordNet [73] and Probase [74]. The abstract knowledge is then manually verified to form a new commonsense knowledge base AbstractATOMIC [71] and AbsPyramid [72].

Besides, recent years researchers start to focus on annotating commonsense in downstream applications, such as grounding to narratives or dialogues [75, 76, 77], persona attribution [78, 70], using commonsense for real-world applications [79].

Visual/Multimodal Commonsense In addition to the commonsense knowledge represented in natural language, commonsense also exists in the vision domain. For example, the relations between visual objects in the real world can be grounded in the commonsense relations between concepts. Specifically, Visual Relationship Detection (VRD) [80] is proposed and annotated to study the relations between pairs of objects in images, such as "person riding a bicycle" or "dog chasing a ball." VisualGenome [81] is later proposed, which annotates over 108,000 images to study the language descriptions of objects, attributes, and relations, to improve computer models' performance on cognitive tasks such as image description and question answering. Besides those grounded concept-concept relations, other resources annotate visual abductive commonsense reasoning signals [82], inferential knowledge of actions, states, and events in images [28], multimodal script knowledge in videos [83], and so on.

Despite their high quality, such CSKBs suffer from limited scale and coverage over various entities and eventualities, as well as potential human bias [84] possibly including concept bias. This leads to other types of commonsense acquisition to improve coverage and potentially alleviate bias.

2.1.2 Information Extraction

To scale up the size of CSKBs, automatic extraction from large corpora based on dedicated schema and templates comes out as an alternative solution [85, 86, 87]. The knowledge base constructed using Information Extraction (IE) can be classified into three categories based on the types of the element of knowledge, namely entity-based, eventuality-based, and statement-based.

Regarding entity-based knowledge bases, KNEXT is an early attempt to mine typical commonsense propositions from corpora [85], while the renowned ConceptNet is originally extracted upon the annotations from OMCS [88]. WebChild [86, 87] extracted relations among activities from narrative texts using semi-supervised label propagation via graph constructed from WordNet and Web data. Regarding eventuality-based knowledge bases, Knowlywood [87] uses semantic parsing tools to extract verb-(object-)based events from TV scenes and novels to build an event knowledge graph. VoCSK [89] proposes a taxonomy-guided induction method for automatically acquiring implicit verb-oriented commonsense knowledge from verb phrases, employing an entropy-based filter to reduce noise and a joint model combining the minimum description length principle with a neural language model to generate concept-level knowledge. ASER [90, 91] proposes a unique paradigm to retrieve a large-scale CSKB for eventualities as linguistic graphs of predicates and their arguments linked by various discourse relations. Regarding statement-based knowledge bases, Ascent [92] and Ascent++ [93] present an advanced methodology for creating a large-scale commonsense knowledge base composed of expressive statements rather than simple triples, encompassing composite concepts and refined assertions to achieve superior precision and recall over existing CSK collections, with proven effectiveness in supporting QA tasks.

However, without human supervision, such methods suffer from much lower quality and larger noise despite much their larger scale. Moreover, they are particularly exposed to reporting bias, including concept bias. Hence such CSKBs alone may not be ideal for our goal to acquire high-quality knowledge covering various entities and eventualities.

2.1.3 Language Model as Knowledge Bases

Another way of acquiring commonsense knowledge is to generate it directly from language models. With knowledge from enormous pretraining data internalized into language models, a pile of works is conducted to mine knowledge from them [39, 94, 95, 96]. One can simply prompt a vanilla language model with self-defined templates or adopt various prompt discovery methods [97, 98, 99]. With the development of the capacity of language models, current backbones such as GPT3 [100], ChatGPT [41], and GPT4 [101] are able to generate commonsense knowledge with even higher accuracy than humans. Relevant works include ATOMIC-10x [102], which prompts GPT3 with several in-context exemplars, coupled with a critic filter to determine the final plausibility. NovaCOMET [103] leverages an auditable discrete knowledge graph, NOVATOMIC, which is constructed using a similar way of text generation as in ATOMIC-10x but using fewer examples. A T5 [38] 11B model is used as the backbone to fine-tune a better commonsense generator.

Furthermore, they are directly applicable to commonsense reasoning tasks [104]. However, language models merely based on induction from word co-occurrence may not generalize well to diverse entities and eventualities, which is why we choose to use information extraction to acquire candidates first and then use language models to discriminate them.

2.2 Commonsense Reasoning

Commonsense reasoning in artificial intelligence (AI) refers to the ability to make assumptions about the nature and characteristics of everyday situations that humans encounter, similar to how humans would do it. In NLP, several popular commonsense reasoning tasks exist, including reasoning over CSKBs, commonsense question answering, and commonsense reasoning in downstream scenarios such as dialogue, narrative, and solving real-world problems.

2.2.1 Commonsense Knowledge Graph/Base Reasoning

Regarding conventional knowledge bases like Wordnet [34] and Freebases [105], tasks involving completion and population have been well-studied as transductive and inductive link prediction problems in the field of graph neural network [106, 107, 108, 109, 110]. Methods powered by pre-trained language models have also been studied in these tasks thanks to the models' representation power [111]. Besides completion tasks on conventional entity-centric KBs like Freebase, completion tasks on CSKBs are also studied on ConceptNet and ATOMIC. Bi-linear models are used to conduct triple classification on ConceptNet [112, 113]. Besides, knowledge base embedding models plus BERT-based graph densifier [2, 114] are used to perform link prediction. Specific to the CSKB Population task, [115] proposed KGBertSAGE, a combination of KG-BERT [111] and GraphSAGE [116]. The model performed better over baselines yet still suffered from the out-of-domain problem.

Another line of commonsense knowledge base reasoning is the commonsense knowledge base population, which is the main focus of this thesis. Specific to CSKB Population task on CKBP v1, [115] proposed KGBertSAGE, a combination of KG-BERT [117] and GraphSAGE [116]. The model showed higher performance over baselines yet still suffered from the out-of-domain problem. The follow-up work PseudoReasoner [118] employs the pseudo-labeling technique to solve that problem. Despite the significant gain in performance, PseudoReasoner is still far from human performance, suggesting that CKBP remains a challenging task in commonsense reasoning.

2.2.2 Commonsense Question Answering

Besides curated commonsense knowledge bases or graphs, a large number of reasoning tasks under commonsense relations are developed among domains like social interaction, spatial and temporal relations, and causes or effects of events. The most popular empirical task is Commonsense Question Answering. Ever since COPA [119], many Commonsense QA datasets have been developed regarding general commonsense in human's daily lives, including CommonsenseQA [120] and CosmosQA [121]. With the trend of investigating social and situational scenarios, SWAG [20], SocialIQa [20], and Dream [122] are proposed as benchmarks concentrated on social environment or human dialogues. Moreover, numerical or physical commonsense datasets like VerbPhysics [123], NumerSense [124], and PhysicalIQA [125] are proposed to test neural models reasoning ability about naive physics and numerical senses.

2.2.3 Real-world Commonsense Reasoning

Besides purely testing the knowledge-understanding ability of AI systems, another line of commonsense reasoning tasks involves studying the commonsense reasoning ability of models in real-world tasks, such as dialogue systems, narrative understanding, and social interactions. For example, ComFact [75] studies the grounding and linking of commonsense knowledge in dialogues and narratives. Peacok [70] studies the persona attributes for improving more engaging and human-like dialogues. Crow [79] studies commonsense in different aspects, such as physical, temporal, and social reasoning.

2.3 Complex Reasoning

2.3.1 Complex Logical Query

Recent years have witnessed significant progress in reasoning on one-hop relational data [106, 126, 127]. In addition to one-hop reasoning, further works have explored handling complex logical structures, involving *reasoning on unobserved edges and multiple entities and variables* [128, 129, 130, 131]. In this thesis, we focus on conjunctive logical queries [63], a subset of first-order logic that is defined with logical operators such as existential quantifiers $\exists$ and conjunctions $\wedge$. Conjunctive logical queries require a set of anchor entities, $\mathcal{V}$, a unique target entity $V_?$ representing the answer to the query, and a set of existential quantified variables $V_1, \cdots, V_m$, and are defined as the conjunction of literals $e_1, \cdots, e_n$:

$$q = V_?, \exists V_1, \cdots, V_m : e_1 \wedge e_2 \wedge \cdots \wedge e_n \tag{2.1}$$

where e_i is an edge involving variable nodes and anchor nodes, satisfying $e_i = r(v_j, V_k), V_k \in \{V_?, V_1, \cdots, V_m\}$, $v_j \in \mathcal{V}, r \in \mathcal{R}$, or $e_i = r(V_j, V_k), V_j, V_k \in \{V_?, V_1, \cdots, V_m\}, j \neq k, r \in \mathcal{R}$. $\mathcal{R}$ is the set of relations defined in the KB.

Previous efforts on answering logical queries on knowledge graphs focus on constructing box embeddings [128], embeddings based on beta distributions [132], particle simulations [133], and computation tree optimization [134]. Other related works focus on leveraging two-hop projection and intersection queries in ConceptNet to improve commonsense question answering [135], inferring missing entities in verbalized complex queries on factual knowledge graphs [136], and developing an LLM agent for complex operators within the KG [137]. Instead of relying on embeddings or limited query types for matching synthetic logical queries, we leverage the concept of logical queries to effectively acquire complex reasoning data from CSKGs with minimum human effort.

2.3.2 Complex Reasoning in Natural Language Processing

There has been a surge of *complex* reasoning tasks in NLP in general [61], including compositional reasoning [138, 139], knowledge retrieval [140, 141], grounding [142], and complex commonsense reasoning such as reasoning on complex narratives or multiple events [8, 9]. In these NLP tasks, the common feature is that they require reasoning over multiple pieces of information. For example, compositional reasoning, such as StrategyQA [138], requires multi-step reasoning based on some retrieved evidence. In knowledge retrieval tasks such as HotpotQA [140], multiple paragraphs from different Wikipedia pages are used as supporting evidence for performing reasoning. In commonsense reasoning, for the single-hop tasks such as CommonsenseQA [120] or SocialIQA [21], even though the knowledge required to solve the problems is usually single-hop, it requires an implicit process of grounding the context to implicit commonsense knowledge, which contributes to the complexity of reasoning.

To tackle complex reasoning, typical methods include knowledge augmentation and chain-of-thought prompting. Knowledge-augmented methods usually involve a module to retrieve relevant knowledge and a module to encode the knowledge to be fused to the main reasoner, e.g., an LLM. Typical methods include KagNet [143], MHGRN [144], QAGNN [145], and GreaseLM [146], which encodes the relevant knowledge retrieved from a knowledge graph and then use Graph Neural Networks to encode the supporting knowledge graph. Another popular line of work is Dense Passage Retrieval (DPR) [147], which retrieves and embeds Wikipedia passages as supporting facts. In the era of Large Language Models, a dominant approach to solving complex tasks is Chain-of-thought (CoT) [148, 149]. The idea is to encourage language models to generate the rationales or reasoning steps implicitly during the prompting of LLMs. This is also considered as an "emergent ability" of language models where CoT can significantly improve the reasoning performance on complex tasks. There are also variants for CoT which include decomposing the reasoning questions to sub-questions (Least2Most [150]), chain-of-though with active learning [151], deductive verification [152], abstraction [153], self-consistency [154], and so on.

CHAPTER 3

MINING LARGE-SCALE COMMONSENSE KNOWLEDGE FROM DISCOURSE KNOWLEDGE

In this chapter, we first study the possibility of transferring linguistic knowledge to *if-then* situational commonsense knowledge.

3.1 Preliminary

Understanding commonsense knowledge has long been one of the ultimate goals of the artificial intelligence field. To achieve that goal, many efforts have been devoted to acquiring commonsense knowledge. For example, ConceptNet [32] (originally known as Open Mind Common Sense (OMCS) [12]) and OpenCyc [65] leverage expert annotation and integration of existing knowledge bases to acquire high-quality commonsense knowledge about human-defined relations. The majority of these relations are factoid commonsense such as *isA*, *partOf*, *attributeOf*, etc. Recently, the focus of sequences of events and the social commonsense relating to them has drawn a lot of attention. ATOMIC [56] is such a knowledge base about inferential knowledge organized as typed *if-then* relations with variables being events and states. Different from traditional knowledge bases, events and states are usually more loosely-structured texts to handle diverse queries of commonsense represented by our natural language. Though being potentially useful for solving commonsense reasoning applications, such kind of commonsense knowledge also brings new challenges for machines to acquire new knowledge of the similar type and make inferences.

3.1.1 Limitations of Current Commonsense Acquisition Methods

First, the knowledge acquired by ATOMIC is based on crowdsourcing, which is relatively more expensive than other automatic information extraction methods. To overcome this

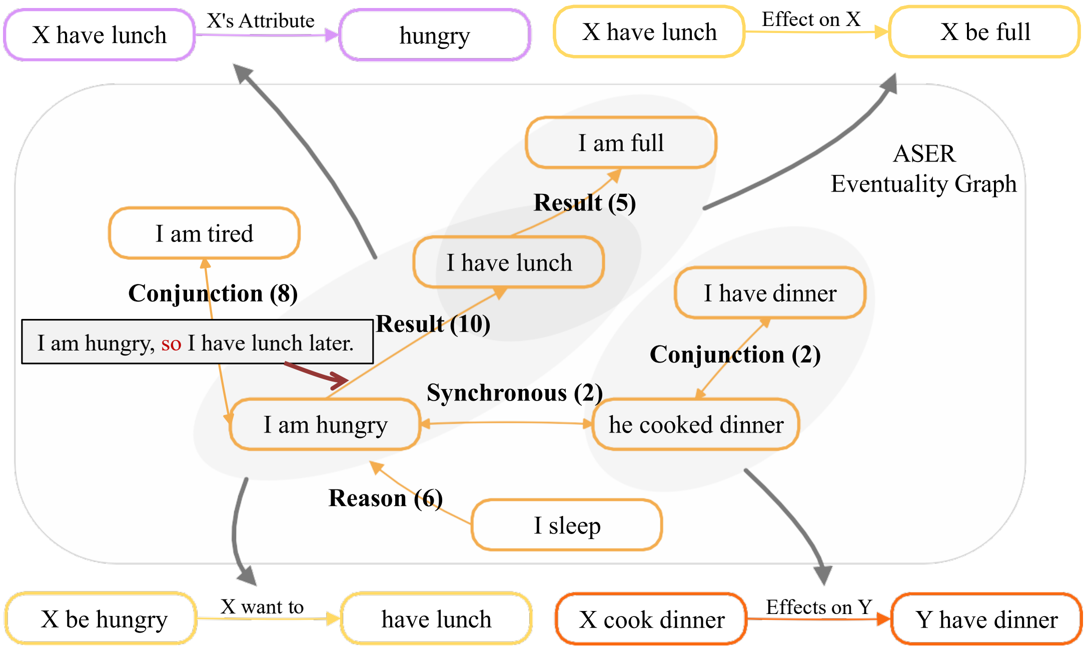


Figure 3.1: An illustration of DISCOS. Eventualities from ASER are connected by directed edges denoting the corresponding discourse relationships. DISCOS aims to transform the discourse edges in ASER to *if-then* commonsense edges. For example, an ASER edge (“I am hungry,” *Result*, “I have lunch”) will be transformed to (*if* “X be hungry,” *then X Want to*, “have lunch”) commonsense tuple. Other discourse edges can also entail other commonsense relations.

problem, COMET [56] is proposed to finetune a large pre-trained language model (i.e., GPT [36]) with existing commonsense knowledge bases (for example, ATOMIC) such that they can automatically generate reasonable commonsense knowledge. Even though COMET can generate high-quality, complex commonsense knowledge with the supervised approach, it tends to fit the training data too well to generate novel concepts. This is usually called a *selection bias* problem in statistical analysis [53, 57].

On the other hand, although information extraction may also be subject to *reporting bias* [52], where the frequencies may not truly reflect the relative likelihood in the real world, it can provide many candidate examples that can be evaluated by a machine learning model trained on human-annotated data. For example, ASER [58] uses frequent syntactical patterns to extract eventualities (such as activities or processes, events, and states) in a dependency parse of a sentence. Then it forms linguistic relations between eventualities based on discourse markers (such as “and,” “but,” etc.) Such an automatic extraction approach can easily scale up to two orders of magnitude larger than human annotations. However, it is not trivial to leverage such a knowledge resource. First, ASER and ATOMIC have different formats. As shown in Figure 3.1, the knowledge in ASER is mostly natural language, for example, “I am hungry,” whereas, in ATOMIC, person entities are mostly aggregated, for example, “Person X be hungry.” Thus, aligning ATOMIC with ASER requires additional efforts to explore both knowledge bases in depth. Second, while some discourse relations extracted in ASER can naturally reflect the *if-then* relations, they are not all valid for each of the *if-then* relations, with variables being events and states. For example, a *Succession* relation in ASER, which is usually extracted by connectives such as “after” and “once,” cannot be used as a candidate relation for the *Stative* relation in ATOMIC because by definition, the *Stative* represents the state of the agent at the same time or before the base event happens, which is opposite from the chronological order of *Succession*.

Last but not least, although it is widely accepted that the graph substructure can be helpful in making predictions and inferences in entity-centric knowledge graphs [60], existing commonsense knowledge-based models [56, 40] still treat the prediction as a translational problem for the triplets in the knowledge base and do not consider the subgraph structures. It is also not trivial to leverage graph structures in commonsense knowledge acquisition. First, there is no existing graph structure in ATOMIC, as the labeling procedure only considers the head, the tail, and their relations. There are few overlaps between heads and tails given that both can be arbitrary texts. The heads and tails can form a bipartite graph, but graph convolution in such a graph may not provide additional information compared to direct representation learning for nodes because tails can be conditionally independent given a head. However, with ASER, which is a more structural knowledge graph, it is possible to perform more complicated reasoning over the substructures. Second, as we mentioned that heads and tails are loosely-structured texts in both ATOMIC and ASER, a contextualized representation model should be applied to them for better representations. As a result, when developing a graph-based model for commonsense acquisition, both the scalability and effectiveness should be carefully considered.

To address the above challenges, in this section, we propose a new commonsense knowledge acquisition framework, DISCOS (from DIScourse knowledge to COmmon-Sense knowledge), which leverages the large-scale eventuality-centric discourse knowledge in ASER to enrich the inferential commonsense knowledge in ATOMIC. Figure 3.1 shows an example of the results. Different from existing mechanisms such as tail node prediction adopted in COMET [56] and link prediction in knowledge base completion tasks used by KG-Bert [111], we propose a knowledge base population approach for DISCOS. This can be done by first mapping ATOMIC nodes to ASER nodes, and then performing a transductive learning algorithm which is based on both contextualized text representation (i.e., BERT [35]) and a graph-related representation (i.e., graph neural networks [116]) to aggregate neighborhood information to jointly make decisions on whether we can populate the ATOMIC relations to a pair of ASER nodes. Experiments demonstrate that the proposed model inherits the advantage of both text and graph representation learning models. Compared with the learning method trained on ATOMIC only, we significantly improve the novelty and diversity of the acquired commonsense knowledge, with comparable accuracy. Extensive analyses are conducted to analyze the strengths and limitations of DISCOS.

3.1.2 Task Definition

The task of acquiring commonsense knowledge from discourse knowledge graphs is defined as a **Commonsense Knowledge Base Population (CKBP)** task. The seed commonsense knowledge base is denoted as $\mathcal{C} = \{(h, r, t)|h \in \mathcal{H}, r \in \mathcal{R}, t \in \mathcal{T}\}$, where $\mathcal{H}$, $\mathcal{R}$, and $\mathcal{T}$ are the set of the heads, relations, and tails, respectively. Suppose we have another much larger knowledge graph extracted from texts via discourse relations, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of all vertices and $\mathcal{E}$ is the set of edges, storing the discourse relations among eventualities.

The CKBP model is trained using a link prediction task over the aligned graph $\mathcal{G}^c$ that contains both the edges from $\mathcal{G}$ and $\mathcal{C}$. The ground truth edges are the corresponding edges from the source commonsense knowledge base $\mathcal{C}$. After learning the edge information from $\mathcal{C}$, in the inference process, the model is asked to predict plausible tail t given head h and relation r as input. Specifically, there are two settings for the inference process: (1) *Existing*

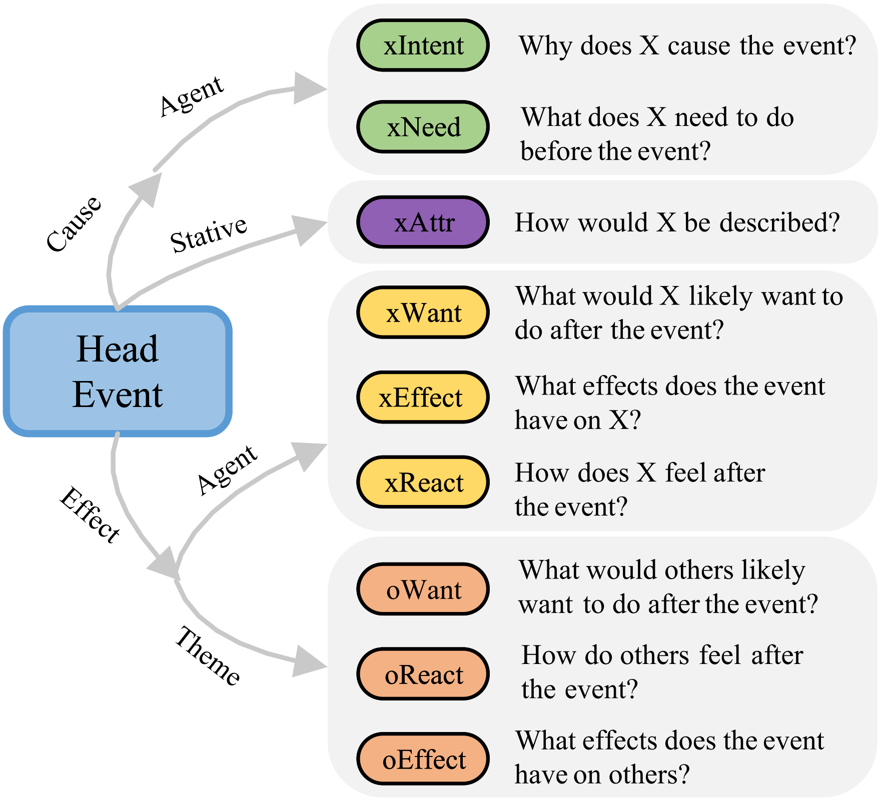

Head Event
Cause
Agent
Stative
Effect
Agent
Theme
xIntent
Why does X cause the event?
xNeed
What does X need to do before the event?
xAttr
How would X be described?
xWant
What would X likely want to do after the event?
xEffect
What effects does the event have on X?
xReact
How does X feel after the event?
oWant
What would others likely want to do after the event?
oReact
How do others feel after the event?
oEffect
What effects does the event have on others?

Figure 3.2: ATOMIC relation definition. The relations are categorized based on chronological order and the subject of events. (1) *cause_agent*: What causes the agent (X) to do the events. (2) *stative*: What is the state of the agent (X). (3) *effect_agent*: What are the effects on the agent (X). (4) *effect_theme*: What are the effects on the theme (Others).

head: Predict tails given head h from the original $\mathcal{C}$, (2) *Novel head*: Predict tails given head h from $\mathcal{G}$ that does not appear in $\mathcal{C}$. While previous works [55, 56] adopt the first setting, we argue that the second setting can generate commonsense knowledge in a much larger scale, considering that $\mathcal{G}$ is much larger.

3.1.3 ATOMIC

We adopt ATOMIC [55] as the seed commonsense knowledge $\mathcal{C}$. ATOMIC consists of 880K tuples across nine relations about day-to-day *if-then* commonsense knowledge (for example, *if* X feels hungry, *then* X wants to have lunch.) Different from structured or canonical knowledge bases, the nodes in ATOMIC are in the form of free-text, which is more expressive in representing everyday commonsense but also makes the matching and generation harder. As shown in Figure 3.2, the nine relation types span over four categories, which are classified based on the order of time and the subject of the events. Detailed illustrations can be found in Figure 3.2.

3.1.4 ASER

ASER [58], a large-scale eventuality-centric knowledge graph that provides explicit discourse relationships between eventualities, is used as the source of discourse knowledge graph $\mathcal{G}$. We use the core part of ASER, which consists of 15 discourse relation types and 10M discourse edges among 27M eventualities. As illustrated in Figure 3.1, the discourse relation ("I am hungry," *Result*, "I have lunch") can be potentially transformed to *if-then* commonsense knowledge, i.e., ("X be hungry," *X want to*, "have lunch.")

Our contributions can be summarized as follows.

• We formulate commonsense acquisition as a Commonsense Knowledge Base Population (CKBP) task, and propose a novel framework, DISCOS, to populate the inferential *if-then* commonsense knowledge in ATOMIC to an eventuality-centric discourse knowledge graph ASER.

• In DISCOS, we develop a model named BERTSAGE to jointly leverage the textual representation and graph representation to discriminate commonsense knowledge. This model can be used as a general approach for commonsense knowledge base population.

• We not only systematically evaluate our framework with commonly used evaluation metrics such as novelty and accuracy using both benchmark dataset and human evaluations, but also thoroughly analyze our models and results as well as the patterns shown in both ATOMIC and ASER to demonstrate that incorporating information extraction results in ASER to enrich the *if-then* relations can indeed provide larger-scale qualified commonsense knowledge.

3.2 The DISCOS Acquisition Pipeline

The overall framework of DISCOS is shown in Figure 3.3. First, the subjects of events in ATOMIC and ASER are quite different; where in ATOMIC the subjects are placeholders

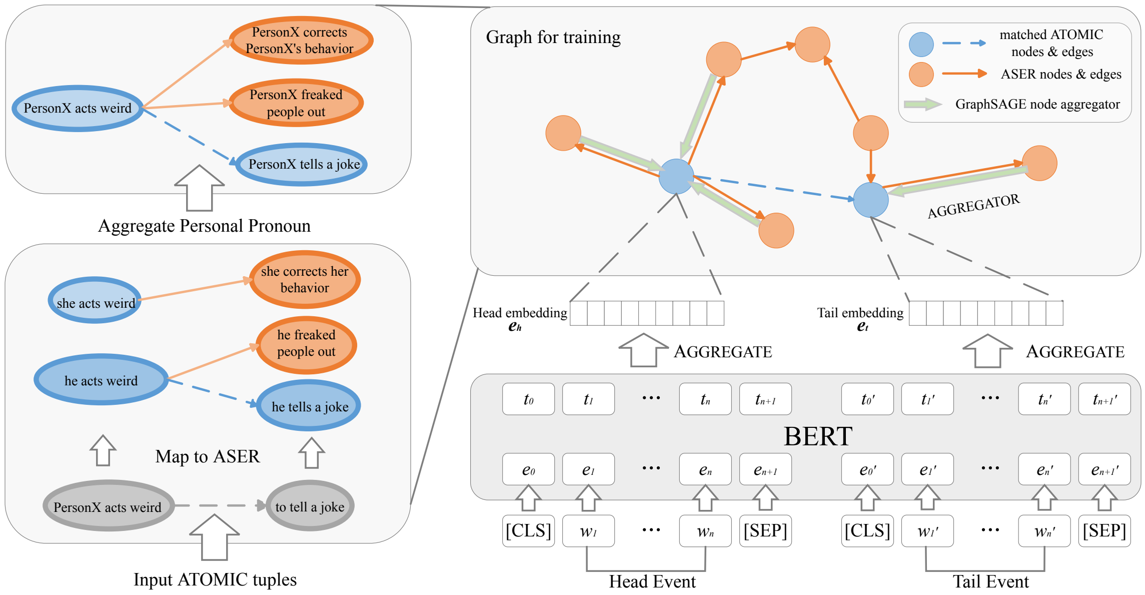

PersonX corrects PersonX's behavior
PersonX acts weird
PersonX freaked people out
PersonX tells a joke
Aggregate Personal Pronoun
she corrects her behavior
she acts weird
he freaked people out
he acts weird
he tells a joke
Map to ASER
PersonX acts weird
to tell a joke
Input ATOMIC tuples
Graph for training
matched ATOMIC nodes & edges
ASER nodes & edges
GraphSAGE node aggregator
AGGREGATOR
Head embedding e_h
Tail embedding e_t
AGGREGATE
AGGREGATE
t_0 t_1 ··· t_n t_{n+1} t_0' t_1' ··· t_n' t_{n+1}'
BERT
e_0 e_1 ··· e_n e_{n+1} e_0' e_1' ··· e_n' e_{n+1}'
[CLS] w_1 ··· w_n [SEP] [CLS] w_1' ··· w_n' [SEP]
Head Event
Tail Event

Figure 3.3: DISCOS overview. First, ATOMIC tuples are mapped to ASER format to acquire candidate commonsense knowledge neighbors from discourse edges in ASER. Then BERTSAGE is used to discriminate whether a (h, r, t) tuple is plausible or not.

like "*PersonX*" and "*PersonY*," while in ASER they are concrete personal pronouns like "she" and "he." So, in order to align the two resources to perform Commonsense Knowledge Braph Population, we first map all heads and tails in $\mathcal{C}$ (ATOMIC) into $\mathcal{G}$ (ASER). Formally, we need a mapping function $M(s)$ to map the input string s into the same format of nodes in $\mathcal{G}$, such that we can find as many $(h, r, t) \in \mathcal{C}$ tuples as possible that can be matched to $\mathcal{G}$ using $M(h)$ and $M(t)$ operations. Next, we leverage a rule $D(v, r), v \in \mathcal{V}, r \in \mathcal{R}$, to select candidate discourse edges in $\mathcal{G}$, given a node $v = M(h), h \in \mathcal{H}$ and a commonsense relation r. After finding all candidate discourse edges under relation r, denoted as $\mathcal{L}(r) = \{(u, v)|(u, v) \in \mathcal{E}\}$, we employ a novel commonsense knowledge population model, BERTSAGE, to score the plausibility of the candidate commonsense tuple (v, r, u). This framework is not restricted to the resource of ATOMIC and ASER but can be well generalized to other resources, as one can change the mapping rules accordingly and use the BERTSAGE model flexibly. Details about each step are introduced as follows.

3.2.1 Aligning ATOMIC and ASER

In ATOMIC, the nodes are eventualities with "*PersonX*" and "*PersonY*" as subjects or objects. However, in ASER, the corresponding eventualities are nodes with concrete personal

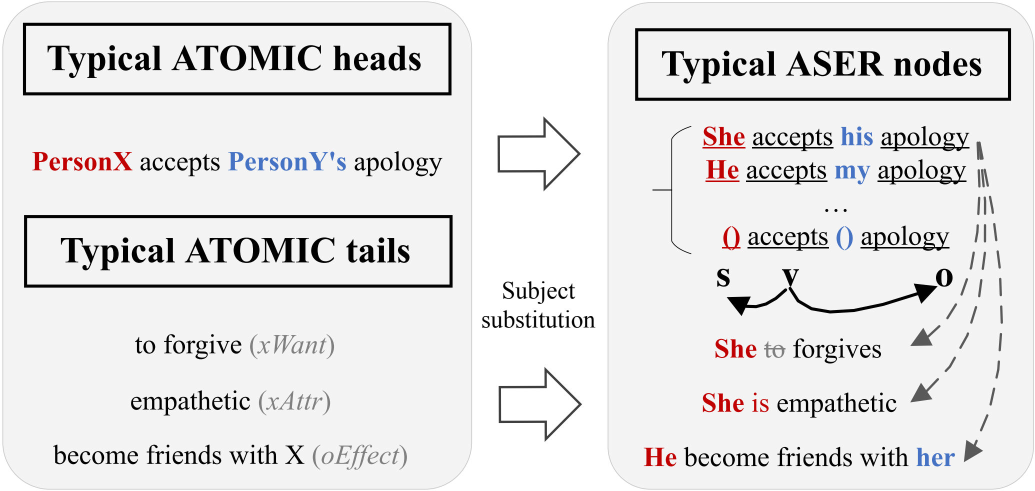


Figure 3.4: An illustration of the alignment between ATOMIC and ASER. We replace the placeholders "*PersonX*" and "*PersonY*" with concrete singular personal pronouns, and add subjects to ATOMIC tails to make them complete sentences.

pronouns, for example, *I*, *she*, *Alex*, and *Bob*. In addition, as the tails in ATOMIC are written by human annotators, the formats can be arbitrary, and sometimes subjects are missing from tails. To effectively align the information in ATOMIC and ASER, based on the above observations, we propose best-effort rules to convert ATOMIC nodes into the format of ASER, as shown in Table 3.1. Examples of the mapping process are shown in Figure 3.4. After conducting the string substitution operations, we use the parser in ASER to parse the acquired text into standard ASER format.

The mapping statistics are shown in Table 3.2, where the average percentage of ATOMIC nodes that can be detected in ASER, denoted as coverage, is 62.9%. It is worth noting that the relation with the highest coverage is *xAttr*, where the tails are mostly adjectives. By adding a personal pronoun and a *be* in front of the *xAttr* tail, we can find most *stative* eventualities in ASER.

We further study the dependency pattern distribution of ATOMIC heads. The head events of ATOMIC are extracted from various corpora, including Google Ngrams and Wiktionary idioms. The definitions of events [55] are similar to that in ASER. We examine the coverage of their dependency patterns using the parser defined in ASER. There are 13 eventuality dependency patterns defined in ASER, as suggested in the paper [58], for example, s-v-o, s-v-o-p-o ('v' for normal verbs other than 'be', 'n' for nouns, 'a' for adjectives,

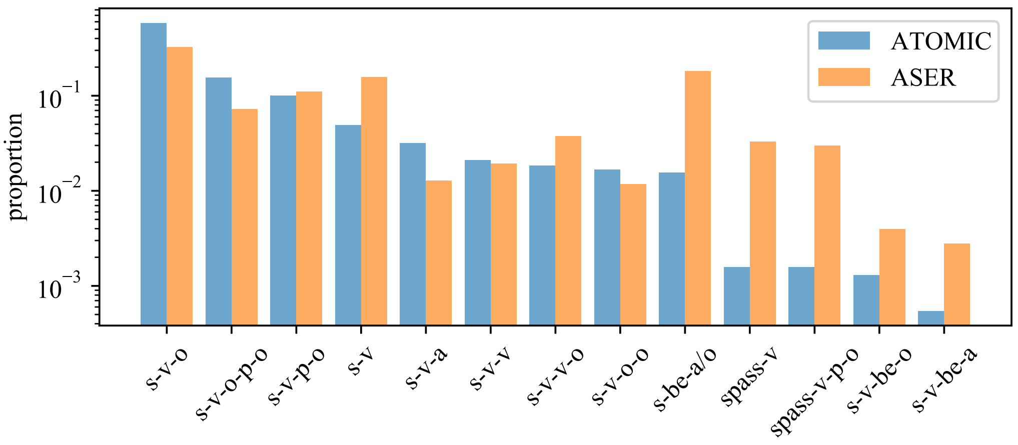


Figure 3.5: Pattern distribution of ATOMIC heads and eventualities in ASER.

and 'p' for prepositions.) The distribution of ATOMIC head patterns and ASER patterns is presented in Figure 3.5. The Pearson r between the distribution of ATOMIC pattern and ASER-core pattern is 0.8136, with $p < 0.01$, showing consistency of ATOMIC and ASER. The syntactical patterns can be used to select eventualities when matching. For example, in "*xAttr*" relation, we restrict the candidate tails in ASER to be of syntactical patterns "s-v-a" and "s-v-o."

3.2.2 Discourse Knowledge Graph Preparation

We then introduce how to select candidate discourse edges from ASER. For a given node u and a relation r, we find the edges based on the rule $D(u, r)$. As we are studying *if-then* relations, the candidate discourse edges in ASER should be consistent with the order of time in the ATOMIC relation r. For example, for a commonsense tuple (h, r, t) in the *effect_agent* category, the event t is an effect of h and thus t should happen at the same time or after the event h. To retrieve ASER discourse edges with the same temporal logic, we first reconstruct an ASER subgraph by selecting specific edge types based on an ATOMIC relation r with rules illustrated in Figure 3.6.

We use the *effect_agent* category as an example. For a given node $u \in \mathcal{V}$, we select the directed (u, v) pairs from ASER, such that there exists either an edge $(u, v) \in \mathcal{E}$ where the edge types are among discourse relations *Precedence* and *Result*, an edge $(v, u) \in \mathcal{E}$ where the edge types are among *Succession*, *Condition*, and *Reason*, or there exists an

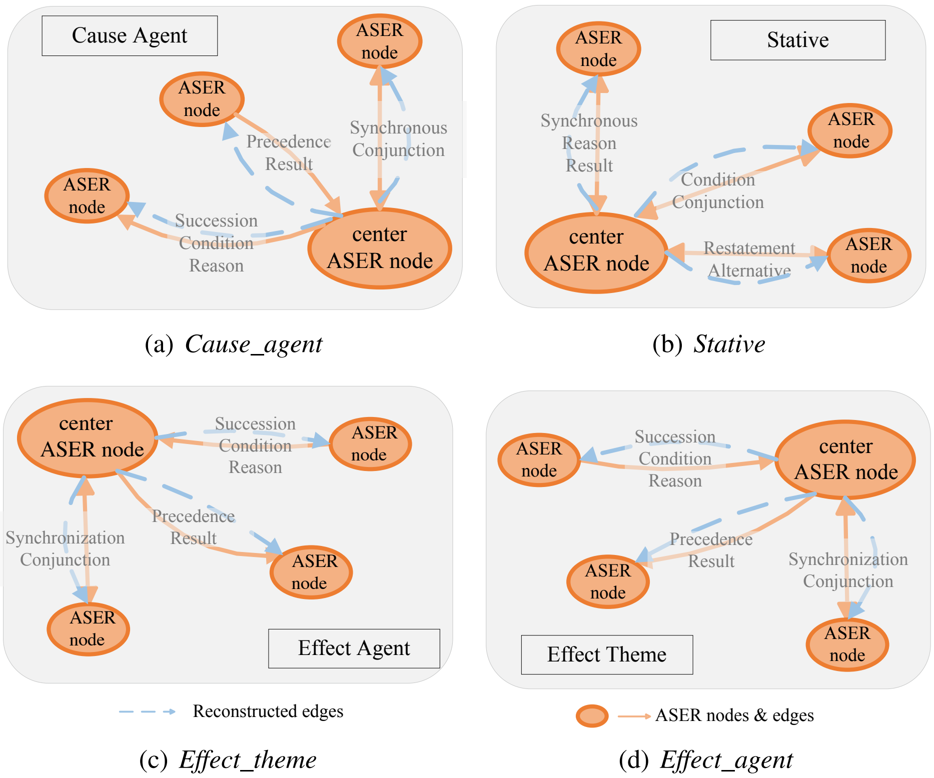
		Mapping rules
Head		Replace *PersonX* and *PersonY* with I/he/she/man/women/person
Tail	xWant/oWant/ xIntent/xNeed	Add a personal pronoun in front of the tail and remove the initial “to”
	xEffect/oEffect	Add a personal pronoun in front of the tail
	xReact/oReact	Add a personal pronoun and “be” in front of the tail
	xAttr	Add a personal pronoun and “be” in front of the tail

Table 3.1: Mapping rules from ATOMIC to ASER.

$e \in \{(u, v), (v, u)\}$ such that the edge types of e is among *Synchronization* and *Conjunction*. In this way, all the selected directed tuples (u, v) represent the same temporal order as in the ATOMIC relation r.

In the next step, we need to distinguish the *theme* categories from *agent* categories. For relations under *effect_theme*, only eventuality pairs (u, v) with different personal pronouns are selected as candidate knowledge, while for other *agent*-based categories, we select eventuality pairs with the same personal pronouns. After this process, combined with all the mapped ATOMIC nodes, we collect all selected edges from $\mathcal{G}$, to form an ASER-induced directed graph $G_r = (V_r, E_r)$ for each relation. V_r is the set of vertices that includes both vertices from $\mathcal{G}$ and the aligned version of $\mathcal{C}$, and E_r is the set of reconstructed edges according to the discourse knowledge selection rules defined above. Here, an edge $(u, v) \in E_r$ can be viewed as a candidate “*if* u, *then* v” relation under r.

After that, we aggregate the nodes in G_r by conducting personal pronoun substitution. For the *agent*-based relations, considering an edge $(u_r, v_r) \in E_r$, we replace the common personal pronouns in u_r and v_r as “*PersonX*,” to be consistent with the ATOMIC format. For other personal pronouns, we map them to “*PersonY*” and “*PersonZ*” according to the order of their occurrences. For the *theme*-based relations, we replace the subject of u_r with “*PersonX*” and v_r with “*PersonY*.” After the personal pronoun substitution operation, we can acquire a unified discourse knowledge graph $G_r^c = (V_r^c, E_r^c)$ in the same format as ATOMIC. The corresponding statistics of all G_r^c are shown in Table 3.2.



(a) *Cause_agent*

(b) *Stative*

(c) *Effect_theme*

(d) *Effect_agent*

Figure 3.6: Discourse knowledge extraction rules for different relation categories. The coral edges represent candidate ASER directed edges to be selected for a certain relation category. The dotted blue edges represent the reconstructed edges in G_r.

3.2.3 Commonsense Knowledge Base Population with BERTSAGE

In our framework, we train a CKBP model on the aligned graph G_r^c.

The basic goal of each step in CKBP is to classify whether a candidate discourse knowledge tuple $(u, v) \in E_r^c$ is a plausible *if-then* commonsense knowledge under relation r. We use the commonsense tuples provided by ATOMIC as the seed ground truth edges. For the negative examples, we explore several different sampling strategies:

- **RAND** (**RANDOM**): Randomly sample two nodes (u, v) from G_r^c such that $(u, v) \notin E_r^c$.

- **O** (**O**thers): Randomly sample two nodes (u, v) from other relations such that $(u, v) \in E_{r'}^c, r' \in \mathcal{R}, r' \neq r$. These negative samples will help the model to distinguish different commonsense relations.

- **I** (**I**nversion): Randomly sample a tuple $(u, v) \in E_r^c$ and add the inversion (v, u) as negative samples. This is used to help the model understand the causal *if-then* relationships, when the input tuples have similar semantic meanings

		ATOMIC		ASER G_r^c	
	Coverage($\%$)	#nodes	#edges	#nodes	#edges
oEffect	31.1	25,328	57,801	170,086	381,135
oReact	87.3	22,970	59,839	95,169	320,543
oWant	61.6	38,892	107,588	177,057	424,409
xAttr	95.8	32,959	174,429	167,869	698,785
xEffect	33.1	43,840	78,644	217,416	721,079
xIntent	33.8	33,789	46,789	179,665	625,144
xNeed	52.9	51,206	92,428	207,317	698,770
xReact	88.7	32,670	99,162	145,216	528,918
xWant	58.8	61,149	114,217	220,786	724,546
Head	56.3	-	-	-	-
Average	62.9	38,089	92,322	175,620	569,259

Table 3.2: Mapping statistics. The ATOMIC columns show the nodes and edges statistics of the graph produced by tuples in ATOMIC. The ASER G_r^c column shows the statistics of the ASER-induced graph for a relation r after personal pronoun aggregation.

• **S** (**S**huffling ATOMIC): Randomly select u from the set of ATOMIC heads, and v from the set of ATOMIC tails under relation r. Add a negative sample if (u, v) is not connected by an existing ATOMIC edge. This mechanism will prevent the model from assigning high scores only to nodes that have appeared in the ATOMIC training set.

To effectively encode both the semantic meaning of eventuality nodes and their neighbors on the overall graph, as shown in the right part of Figure 3.3, we propose a model BERTSAGE that contains two components: (1) a node encoder based on BERT that embeds the semantic meaning of nodes; (2) a graph encoder that learns and aggregates relational information from the discourse graph. The details are as follows.

• **Node encoder:** We use the pre-trained language representation mode BERT [35] to encode all the nodes in the dataset. For a node $v = [w_1, w_2, \cdots, w_n]$ with n word tokens, we add a [CLS] token in the beginning of each sentence as w_0 and a [SEP] token at the end of it as w_{n+1}. We denote the contextualized representation provided by BERT as $[\boldsymbol{e}_{w_0}, \boldsymbol{e}_{w_1}, \cdots, \boldsymbol{e}_{w_{n+1}}]$, $\boldsymbol{e}_{w_i} \in \mathbb{R}^d$, where d is the dimension of BERT embeddings, $\boldsymbol{e}_{w_0}$ and $\boldsymbol{e}_{w_{n+1}}$ are the embedding of [CLS] and [SEP] tokens, respectively. We then use the average pooling to acquire the final node representation as $\boldsymbol{e}_v = \sum_{i=0}^{n+1} \boldsymbol{e}_{w_i}/(n+2)$.

• **Graph encoder:** To effectively encode the semantics from neighbor events on the discourse graphs, we propose to use GraphSAGE [116] to aggregate the neighbor information

Model	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant
BERT	90.60	97.05	93.95	96.21	87.85	89.69	89.93	93.96	89.73
BERTSAGE	**91.10***	**97.29**	**94.21**	**96.33**	**89.49***	**90.48***	**91.10***	**94.02**	**90.91***

Table 3.3: Evaluations on the CKBP link prediction experiments. We report the accuracy in test set here as the number of positive and negative samples are balanced. * after bold figures indicates that the improvement of BERTSAGE model is significant with z-test $p < 0.05$.

of a given node v.

Given a node v, we first acquire its contextualized representation $\boldsymbol{e}_v$, and then calculate the embeddings of v's neighbors in G_r^c, which are denoted as $\mathcal{N}(v)$. Here, $\mathcal{N}(v)$ is a fixed size neighbor set uniformly sampled from all the neighbors of v. The hidden representation after the GraphSAGE layer $\boldsymbol{h}_v$ is computed as follows:

$$\boldsymbol{h}_{\mathcal{N}(v)} \leftarrow \text{AGGREGATE}(\{\boldsymbol{e}_u, \forall u \in \mathcal{N}(v)\}), \tag{3.1}$$

$$\boldsymbol{h}_v \leftarrow \sigma(\boldsymbol{W} \cdot \text{CONCAT}(\boldsymbol{h}_v, \boldsymbol{h}_{\mathcal{N}(v)})). \tag{3.2}$$

• **Output layer:** For an input candidate tuple $(u, v) \in G_r^c$, on top of the overall representation given by BERTSAGE $[\boldsymbol{h}_u, \boldsymbol{h}_v]$, we apply an output layer $f_r(u, v) = \text{Softmax}([\boldsymbol{h}_u, \boldsymbol{h}_v]\boldsymbol{W}'^{\top} + \boldsymbol{b})$, $\boldsymbol{W}' \in \mathbb{R}^{2\times d}, \boldsymbol{b} \in \mathbb{R}^2$ to make the final prediction.

3.3 Experiments

We introduce the experimental settings and results of all the experiments in this section. Both learning and inference processes in Section 3.1.2 are studied here. For the learning part, the task is a link prediction task, and thus, we evaluate the performance automatically using accuracy based on the existing annotated commonsense knowledge as positive examples and automatically sampled edges as negative examples. For the inference part, as the goal to acquire novel commonsense knowledge is similar with previous works in ATOMIC [55] and COMET [56], we adopt human evaluation to evaluate the quality of the newly acquired knowledge, and then use novelty and diversity as additional evaluation metrics accordingly.

3.3.1 Learning for CKBP

Settings We first train the BERTSAGE model for Commonsense Knowledge Base Population. We evaluate the performance of link prediction using accuracy. We use the edges derived from ATOMIC as positive examples. 20% of the negative examples are randomly sampled using **O**, 10% of them using **I**, and the rest using **RAND**, as defined in Section 3.2.3. Detailed ablation studies about negative sampling techniques are presented in Section 3.4.1. Considering that G_r^c is much larger than ATOMIC, we restrict the size of G_r^c in the following ways. (1) We only select the subgraph of G_r, which is induced by the one-hop neighbors of all the ATOMIC nodes. (2) For the subgraph acquired in the first step, we add two-hop neighbors into the graph for the nodes whose degrees are less than a threshold k. k is set to 20 in practice. We use *bert-base-uncased*[1] [35] as the encoding layer for the classification model and the dimension of the hidden embeddings is 768. For the neighbor function $\mathcal{N}(u)$, we set the neighbor size to be 4. The batch size is set to 64. We use the train/dev/test split defined in ATOMIC [55]. To clearly show the contribution of the proposed BERTSAGE model, we compare it with a modified version of KG-Bert [111], denoted as BERT baseline for short, on the link prediction task. The only difference between BERTSAGE and BERT is that we incorporate the semantics about neighboring events to get the final representation in BERTSAGE. We use the same setting for BERT as defined above, except for the graph module.

Result We report the accuracy of the CKBP models on the test set with prepared negative edges. From the results in Table 3.3, we can see that adding a GraphSAGE layer over the BERT baseline will improve the classification results on all relation types. These results prove our assumptions that adding information about the neighbor events on the discourse graph can help generate better event representations. Among nine relations, the improvements are significant with z-test $p < 0.05$ on five types. One interesting finding is that this improvement is in positive correlation with the graph complexity in Table 3.2. In general, GraphSAGE will contribute more to the performance when the graph is more complex.

[1] https://github.com/huggingface/transformers

	dist-1		dist-2	
relations	COMET	DISCOS	COMET	DISCOS
oEffect	60.3	**66.7**	76.3	**89.3**
oReact	**35.5**	33.5	13.5	**35.9**
oWant	46.6	**69.0**	84.1	**93.8**
xAttr	8.3	**26.0**	4.2	**27.4**
xEffect	58.4	**67.2**	81.8	**90.4**
xIntent	42.9	**61.5**	75.7	**87.3**
xNeed	41.4	**63.6**	75.7	**88.4**
xReact	27.1	**29.3**	12.1	**32.9**
xWant	42.2	**65.3**	78.7	**91.5**
Average	38.3	**52.9**	55.0	**70.0**

Table 3.4: Diversity grouped by all the relations for the *existing head* setting in the inference process. We report the diversity of the top 10 generations or retrieval of COMET and DISCOS. Dist-k indicates the proportion of unique k-grams.

3.3.2 Inference for CKBP

Settings We evaluate the capability of the above BERTSAGE model with the inference part for acquiring new commonsense knowledge in CKBP. The goal of the inference part is similar to that in COMET. As there is no ground truth for the newly generated nodes or edges, we conduct human evaluation for the quality. Besides accuracy, we also use automatic metrics related to novelty and diversity to demonstrate the properties of the acquired commonsense knowledge. While COMET uses a neural generation method to generate tails, in DISCOS we use BERTSAGE to rank the candidates provided in ASER given heads and relations. Similar to COMET, for the *existing head* setting introduced in Section 3.1.2, we propose to evaluate the acquired commonsense knowledge from all three perspectives, i.e., quality, novelty, and diversity. While for the *novel head* setting, as the heads are already novel, we only evaluate the Quality of the retrieved knowledge.

• **Quality**: We evaluate the quality of acquired commonsense knowledge using annotators from Amazon Mechanical Turk (AMT.) For each relation in ATOMIC, we randomly sample 50 head events from the testing set and ask the annotators if they think the generated tuple makes sense. For COMET, we use beam 10 top 10 as the decoding mechanism to generate 10 commonsense knowledge for each head event. For DISCOS, we select the tuples ranked top 10 by the BERTSAGE model.

• **Novelty**: We first evaluate the novelty of acquired commonsense knowledge with two novelty indicators, the proportion of generated tails that are novel (NT_t), and the proportion of novel tails in the set of all the unique generated tails (NU_t.)

• **Diversity**: Last but not least, considering that the novelty is evaluated based on string match, which cannot effectively distinguish whether a system is generating many different novel concepts or just similar but not identical concepts. Following previous works [155, 156], we report diversity indicators dist-1 and dist-2, the proportion of distinct unigrams and bigrams among the total number of generated unigrams and bigrams. We evaluate the diversity of generated knowledge given the same head and relation and calculate the average among all the heads.

For COMET, we use the publicly available official implementation[2]. All the experimental settings are the same as in the original paper. Similar to the decoding mechanisms in the COMET paper, we use beam search top k to retrieve k generated tails.

Result The overall quality[3], novelty, and diversity of COMET and DISCOS are shown in Table 3.7, 3.6, and 3.4, respectively. From the results, we can make the following observations. Based on our crowd-sourcing results, DISCOS can achieve comparable or better quality on *effect_theme* relations (*oEffect*, *oReact*, and *oWant*) and *cause_agent* relations (*xIntent* and *xNeed*) among the nine relations. The results indicate that rich commonsense knowledge is indeed covered by the discourse graph and the proposed DISCOS framework can effectively discover them. At the same time, we also notice that DISCOS can significantly outperform COMET in terms of novelty. For example, for some relations like *xAttr*, *oReact*, and *xReact*, COMET hardly generates novel tails despite increasing the size of beam search while a large portion of the DISCOS knowledge is novel. One reason behind this is that COMET fits the training data too well, and the training set is similar to the test set. As a result, it tends to predict the concepts it has seen in the training set rather than something new. Last but not least, similar to the novelty, DISCOS also outperforms COMET in terms of the diversity, which is mainly due to the limitation of beam search

[2] `https://github.com/atcbosselut/comet-commonsense`

[3] We present the original human annotation results from the ATOMIC paper as a reference. However, as we employ different annotators, they are not comparable with our results.

as it often generates very similar sentences. As DISCOS is a classification model rather than a generation model, it does not suffer from that problem. To conclude, compared with COMET, DISCOS can acquire much more novel and diverse commonsense knowledge with comparable quality.

To further demonstrate that DISCOS has the potential to acquire commonsense knowledge without the help of human-defined heads, we evaluate it with the *novel head* setting. Here, only the relation r is provided, and the model is asked to retrieve the novel (h, t) pairs from ASER. Specifically, we select the tuples scored higher than 0.5 by the BEST-SAGE model and randomly sample 100 tuples from each relation for human evaluation. To make sure the acquired knowledge is not observed by the model, only novel concepts are evaluated.

From the results in Table 3.7, we can see the potential of DISCOS in directly mining high-quality novel commonsense knowledge from the raw graph of ASER. For example, it achieves over 70% accuracy on three relations (*"oEffect"*, *"xEffect"*, and *"xReact"*.) Following this experimental setting, we successfully convert ASER into a large scale commonsense knowledge base DISCOS-ATOMIC, which contains 3.4 million complex commonsense knowledge in the format of ATOMIC, without using any additional annotation effort.

3.4 Analysis

3.4.1 Ablation Study

In this subsection, we will present ablation studies on the effects of different negative sampling strategies on the link prediction part of the CKBP task. We tried to use the aforementioned combinations in Section 3.2.3 to generate the negative examples for both the training and testing set, and present the results of link prediction accuracy[4] on the test set in Table 3.5. Specifically, we tried the following combinations:

1. RAND: All the negative examples are sampled randomly from the whole graph.

[4] We select the *xWant* relation as an example.

Train Test	RAND	O20	O20+I10	O20+I10 +S10
RAND	94.40	93.65	93.50	90.88
O20	87.46	91.16	90.93	89.12
O20+I10	87.16	90.72	90.92	89.17
O20+I10+S10	82.80	86.49	**86.85**	86.53

Table 3.5: Ablation study on different negative sampling methods under *xWant* relation, trained using BERTSAGE model. We report the accuracy of the testing set here using the link prediction task in CKBP.

2. O20: 20% of the negative examples are sampled using the mechanism **O**.

3. O20+I10: 20% of the negative examples are sampled using the mechanism **O** and 10% from the mechanism **I**.

4. O20+I10+S10: 20% of the negative examples are sampled using the mechanism **O**, 10% from the mechanism **I**, and 10% from the mechanism **S**.

We highlight the accuracy ranked highest on O20+I10+S10 test set, the hardest negative example set. From the result, we can see that, even though the RAND achieves comparable performance on the simple test set RAND, it suffers a significant performance drop on the other harder ones. The reason behind is that the randomly selected negative examples can only help the model to distinguish the ATOMIC positive examples rather than distinguish the commonsense. This ablation study also demonstrates the importance of including more diverse negative example generation strategies to cover more signals we want the model to learn. In the end, we choose to use O20+I10 negative sampling for training in our final model.

Model	oEffect		oReact		oWant		xAttr		xEffect		xIntent		xNeed		xReact		xWant	
	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}	N_{to}	N_{uo}
COMET @ 1	0.0	0.0	0.0	0.0	2.2	10.4	0.2	0.7	2.0	8.3	4.1	9.9	12.0	29.3	0.0	0.0	9.4	18.1
DISCOS @ 1	**61.2**	**65.0**	**15.5**	**36.3**	**43.2**	**56.7**	**6.8**	**17.2**	**45.3**	**54.2**	**38.5**	**61.6**	**29.7**	**44.0**	**5.5**	**21.4**	**25.6**	**45.8**
COMET @ 2	7.5	23.9	0.0	0.0	3.5	17.8	0.1	0.4	2.3	9.3	6.2	14.8	11.8	32.0	0.0	0.0	10.2	20.0
DISCOS @ 2	**58.4**	**65.9**	**13.7**	**33.8**	**45.1**	**63.0**	**7.1**	**20.1**	**45.3**	**58.6**	**39.5**	**63.0**	**30.5**	**49.3**	**6.2**	**26.5**	**26.3**	**49.1**
COMET @ 5	12.9	30.3	0.1	1.0	7.5	25.4	0.1	0.7	5.3	17.0	8.6	21.4	15.8	35.5	0.1	0.9	11.5	26.5
DISCOS @ 5	**59.9**	**72.8**	**16.5**	**38.8**	**49.9**	**69.6**	**8.9**	**25.8**	**50.2**	**65.3**	**44.9**	**68.8**	**38.2**	**59.5**	**6.9**	**33.7**	**32.2**	**55.1**
COMET @ 10	16.8	40.4	0.4	4.9	9.8	32.4	0.1	0.8	8.0	24.1	12.7	31.2	18.6	41.0	0.4	4.7	12.3	30.5
DISCOS @ 10	**62.9**	**76.2**	**22.5**	**50.4**	**55.8**	**75.4**	**12.0**	**30.4**	**54.5**	**71.1**	**51.7**	**74.1**	**44.2**	**66.2**	**9.1**	**42.8**	**38.1**	**62.0**

Table 3.6: Novelty on the test set grouped by different relations, under *existing head* setting of the inference process. @ k means we evaluate the top k generation or retrieval for a given head h. All improvements by DISCOS are significant with z-test p<0.05.

	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant
DISCOS (novel heads and tails)	70.2	63.2	59.4	69.2	78.2	65.8	67.8	80.0	49.2

Table 3.7: Human annotation on quality for the *novel head* setting (given r to retrieve plausible (h, t) pairs in DISCOS.)

CHAPTER 4

BENCHMARKING COMMONSENSE KNOWLEDGE BASE POPULATION

4.1 Preliminary

Commonsense reasoning is one of the core problems in the field of artificial intelligence. Throughout the development in computational commonsense, commonsense knowledge bases (CSKB) [32, 55] are constructed to enhance models' reasoning ability. As human-annotated CSKBs are far from complete due to the scale of crowd-sourcing, reasoning tasks such as *CSKB completion* [112, 2, 157] and *population* [110] are proposed to enrich the missing facts. The CSKB completion task is defined based on the setting of predicting missing links within the CSKB. On the other hand, the population task grounds commonsense knowledge in CSKBs to large-scale automatically extracted candidates and requires models to determine whether a candidate triple, *(head, relation, tail)*, is plausible or not, based on the information from both the CSKB and the large number of candidates which essentially form a large-scale graph structure. An illustration of the difference between completion and population is shown in Figure 4.1.

There are two advantages of the population task. First, the population can not only add links but also nodes to an existing CSKB, while completion can only add links. The populated CSKB can also help reduce the *selection bias* problem [53] from which most machine learning models would suffer and will benefit a lot of downstream applications such as commonsense generation [56]. Second, commonsense knowledge is usually implicit knowledge that requires multiple-hop reasoning, while current CSKBs lack such complex graph structures. For example, in ATOMIC [55], a human-annotated *if-then* commonsense knowledge base among daily events and (mental) states, the average hops between matched heads and tails in ASER, an automatically extracted knowledge base among activities, states, and events based on discourse relationships, is 2.4 [11]. Evidence in Section 4.2.5 (Table 4.3)

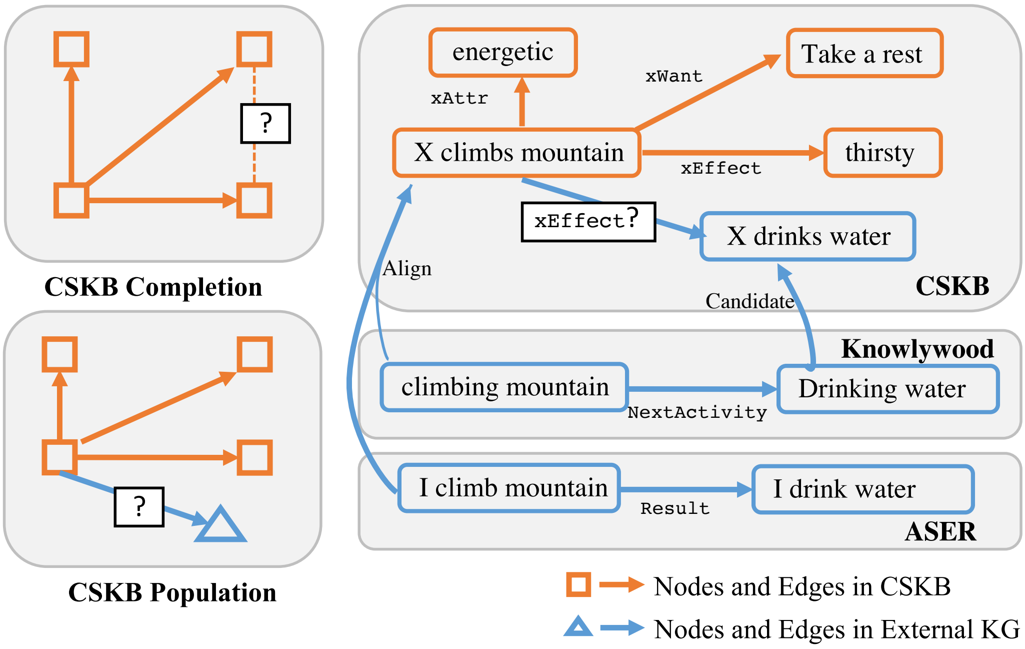


Figure 4.1: Comparison between CSKB completion and population. An example of aligning the eventuality graph with candidate commonsense knowledge triples is also provided.

also shows similar results for other CSKBs. However, reasoning solely on existing CSKBs can be viewed as a simple triple classification task without considering complex graph structure (as shown in Table 4.3, the graphs in CSKBs are much sparser). The population task, which provides a richer graph structure, can explicitly leverage the large-scale corpus to perform commonsense reasoning over multiple hops on the graph.

4.1.1 Limitations of CSKB Population

However, there are two major limitations for the evaluation of the CSKB population task. First, automatic evaluation metrics, which are based on distinguishing ground truth annotations from automatically sampled negative examples (either a random head or a random tail), are not accurate enough. Instead of directly treating the random samples as *negative*, solid human annotations are needed to provide hard labels for commonsense triples. Second, the human evaluation in the original paper of CSKB population [110] cannot be generally used for benchmarking. They first populate the CSKB and then asked human annotators to annotate a small subset to check whether the populated results are accurate or not. A better benchmark should be based on random samples from all candidates, and the scale should be large enough to cover diverse events and states.

To effectively and accurately evaluate CSKB population, in this section, we benchmark CSKB population by firstly proposing a comprehensive dataset aligning four popular CSKBs and a large-scale automatically extracted knowledge graph, and then providing a large-scale human-annotated evaluation set. Four event-centered CSKBs that cover daily events, namely ConceptNet [32] (the event-related relations are selected), ATOMIC [55], ATOMIC$^{20}_{20}$ [10], and GLUCOSE [1], are used to constitute the commonsense relations. We align the CSKBs together into the same format and ground them to a large-scale eventuality (including activity, state, and event) knowledge graph, ASER [58, 11]. Then, instead of annotating every possible node pair in the graph, which takes an infeasible $O(|V|^2)$ amount of annotation, we sample a large subset of candidate edges grounded in ASER to annotate. In total, 31.7K high-quality triples are annotated as the development set and test set.

To evaluate the commonsense reasoning ability of machine learning models based on our benchmark data, we first propose some models that learn to perform CSKB population inductively over the knowledge graph. Then, we conduct extensive evaluations and analysis of the results to demonstrate that the CSKB population is a hard task where models perform poorly on our evaluation set far below human performance.

We summarize the contributions of the section as follows:

(1) We provide a novel benchmark for CSKB population over new assertions that cover four human-annotated CSKBs, with a large-scale human-annotated evaluation set.

(2) We propose a novel inductive commonsense reasoning model that incorporates both semantics and graph structure.

(3) We conduct extensive experiments and evaluations on how different models, commonsense resources for training, and graph structures may influence the commonsense reasoning results.

Glucose	ATOMIC Relations
Dim 1, 6	xEffect, oEffect
Dim 2	xAttr (“feels”), xIntent (otherwise)
Dim 3, 4, 8, 9	Causes
Dim 5, 10	xWant, oWant
Dim 7	xReact, oReact

Table 4.1: The conversion from GLUCOSE relations to ATOMIC$_{20}^{20}$ relations, inherited from [1].

	ATOMIC (No clause)	ATOMIC$_{20}^{20}$ (4 relations)	ConceptNet (Event-centered)	GLUCOSE	# Eventuality
# Triples	449,056	124,935	10,159	117,828	-
Knowlywood	2.63%	2.87%	16.50%	2.96%	929,546
ASER	61.95%	38.50%	44.94%	84.57%	52,940,258

Table 4.2: Overlaps between eventuality graphs and commonsense knowledge graphs. We report the proportion of (h, r, t) triples where both the head and tail can be found in the eventuality graph.

4.1.2 Task Definition

Denote the source CSKB about events as $\mathcal{C} = \{(h, r, t)|h \in \mathcal{H}, r \in \mathcal{R}, t \in \mathcal{T}\}$, where $\mathcal{H}$, $\mathcal{R}$, and $\mathcal{T}$ are the set of the commonsense heads, relations, and tails. Suppose we have another much larger eventuality (including activity, state, and event) knowledge graph extracted from texts, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of all vertices and $\mathcal{E}$ is the set of edges. $\mathcal{G}^c$ is the graph acquired by aligning $\mathcal{C}$ and $\mathcal{G}$ into the same format. The goal of CSKB population is to learn a scoring function given a candidate triple (h, r, t), where plausible commonsense triples should be scored higher. The training of CSKB population can inherit the setting of triple classification, where ground truth examples are from the CSKB $\mathcal{C}$ and negative triples are randomly sampled. In the evaluation phase, the model is required to score the triples from $\mathcal{G}$ that are not included in $\mathcal{C}$ and be compared with human-annotated labels.

4.2 Dataset Preparation

4.2.1 Selection of CSKBs

As we aim to explore commonsense relations among general events, we summarize several criteria for selecting CSKBs. First, the CSKB should be well symbolically structured to be generalizable. While the nodes in CSKB can inevitably be free-text to represent more diverse semantics, we select the knowledge resources where format normalization is conducted. Second, the commonsense relations are encoded as *(head, relation, tail)* triples. To this end, among all CSKB resources, we choose the event-related relations in ConceptNet, ATOMIC, ATOMIC$^{20}_{20}$, and GLUCOSE as the final commonsense resources. For the event-related relations in ConceptNet, the elements are mostly lemmatized *predicate-object* pairs. In ATOMIC and ATOMIC$^{20}_{20}$, the subjects of eventualities are normalized to placeholders "*PersonX*" and "*PersonY*". The nodes in GLUCOSE are also normalized and syntactically parsed manually, where human-related pronouns are written as "*SomeoneA*" or "*SomeoneB*", and object-related pronouns are written as "*SomethingA*". Other commonsense resources like SocialChemistry101 [68] are not selected as they include loosely-structured events.

For ConceptNet, we select the event-related relations `Causes` and `HasSubEvent`, and the triples where nodes are noun phrases are filtered out. For ATOMIC, we restrict the events to simple and explicit events that do not contain wildcards and clauses. As ATOMIC$^{20}_{20}$ itself includes the triples in ATOMIC and ConceptNet, to distinguish different relations, we refer to ATOMIC$^{20}_{20}$ as the new event-related relations annotated in ATOMIC$^{20}_{20}$ [10], which are `xReason`, `HinderedBy`, `isBefore`, and `isAfter`. In the rest of the section, ATOMIC($^{20}_{20}$) means the combination of ATOMIC and the new relations in ATOMIC$^{20}_{20}$.

4.2.2 Alignment of CSKBs

To effectively align the four CSKBs, we propose best-effort rules for aligning the formats for nodes and edges. First, for the nodes in each CSKB, we normalize the *person-centric* subjects and objects as "*PersonX*", "*PersonY*", and "*PersonZ*", etc, according to the order of their occurrence, and the *object-centric* subjects and objects as "*SomethingA*" and "*SomethingB*". Second, to reduce the semantic overlaps of different relations, we aggregate all commonsense relations to the relations defined in ATOMIC$(^{20}_{20})$, as it is comprehensive enough to cover the relations in other resources like GLUCOSE, with some simple alignment in Table 4.1.

ConceptNet. We select `Causes` and `HasSubEvent` from ConceptNet to constitute the event-related relations. As heads and tails in ConceptNet don't contain subjects, we add a "*PersonX*" in front of the original heads and tails to make them complete eventualities.

ATOMIC$(^{20}_{20})$. In ATOMIC and ATOMIC$^{20}_{20}$, heads are structured events with "*PersonX*" as subjects, while tails are human-written free-text where subjects tend to be missing. We add "*PersonX*" for the tails without subjects under *agent*-driven relations, the relations that aim to investigate causes or effects on "*PersonX*" himself, and add "*PersonY*" for the tails missing subjects under *theme*-driven relations, the relations that investigate commonsense causes or effects on other people like "*PersonY*" .

GLUCOSE. For GLUCOSE, we leverage the parsed and structured version in this study. We replace the personal pronouns "*SomeoneA*" and "*SomeoneB*" with "*PersonX*" and "*PersonY*" respectively. For other *object-centric* placeholders like "*Something*", we keep them unchanged. The relations in GLUCOSE are then converted to ATOMIC relations according to the conversion rule in the original paper [1]. Moreover, `gWant`, `gReact`, and `gEffect` are the new relations for the triples in GLUCOSE where the subjects are *object-centric*. The prefix "g" stands for *general*, to be distinguished from "x" (for *PersonX*) and "o" (for *PersonY*).

4.2.3 Selection of the Eventuality KG

Taking scale and the diversity of relationships in the KG into account, we select two automatically extracted eventuality knowledge graphs as candidates for the population task, Knowlywood [158] and ASER [58]. They both have complex graph structures that are suitable for multiple-hop reasoning. We first check how much commonsense knowledge is included in those eventuality graphs to see if it's possible to ground a large proportion of commonsense knowledge triples on the graphs. Best-effort alignment rules are designed to align the formats of CSKBs and eventuality KGs. For Knowlywood, as the patterns are mostly simple *verb-object* pairs, we leverage the *v-o* pairs directly and add a subject in front of the pairs. For ASER, we aggregate the raw personal pronouns like *he* and *she* to normalized "*PersonX*". As ASER adopts more complicated patterns of defining eventualities, a more detailed pre-process of the alignment between ASER and CSKBs will be illustrated in Section 4.2.4. We report the proportion of triples in every CSKB whose head and tail can both be matched to the eventuality graph in Table 4.2. ASER covers a significantly larger proportion of head-tail pairs in the four CSKBs than Knowlywood. The reason behind this is that, on the one hand, ASER is of much larger scale, and on the other hand, ASER contains eventualities with more complicated structures like *s-v-o-p-o* (*s* for *subject*, *v* for *verb*, *o* for *object*, and *p* for *preposition*), compared with the fact that Knowlywood mostly covers *s-v* or *s-v-o* only. In the end, we select ASER as the eventuality graph for population.

4.2.4 Pre-process of the Eventuality Graph

We introduce the normalization process of ASER, which converts its knowledge among everyday eventualities into normalized form to be aligned with the CSKBs as discussed in Section 4.2.2. Each eventuality in ASER has a subject. We consider singular personal pronouns, i.e., "I", "you", "he", "she", "someone", "guy", "man", "woman", "somebody", and replace the concrete personal pronouns in ASER with normalized formats such as "*PersonX*" and "*PersonY*". Specifically, for an original ASER edge where both the head and tail share the same *person-centric* subject, we replace the subject with "*PersonX*" and the subsequent personal pronouns in the two eventualities with "*PersonY*" and "*PersonZ*" according to the order of the occurrence if exists. For the two neighboring eventualities where the subjects are different *person-centric* pronouns, we replace one with "*PersonX*" and the other with "*PersonY*". In addition, to preserve the complex graph structure in ASER, for all the converted edges, we duplicate them by replacing the "*PersonX*" in it with "*PersonY*", and "*PersonY*" with "*PersonX*", to preserve the sub-structure in ASER as much as possible.

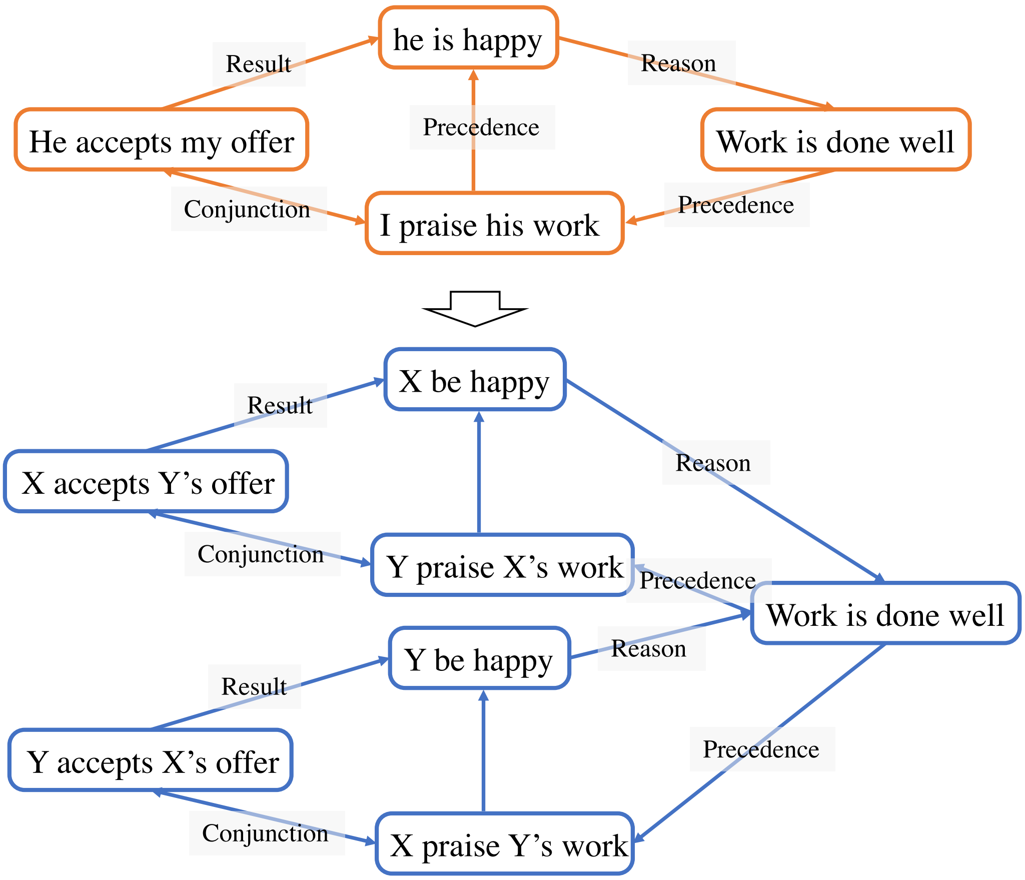


Figure 4.2: An example of normalizing ASER. The coral nodes and edges are raw data from ASER, and the blue ones are the normalized graph by converting “he” and “she” to placeholders “*PersonX*” and “*PersonY*”

An illustration of the converting process is shown in Figure 4.2. The normalized version of ASER is denoted as ASER$_{norm}$.

4.2.5 The Aligned Graph $\mathcal{G}^c$

With the pre-process in Section 4.2.2 and 4.2.4, we can successfully align the CSKBs and ASER together in the same format. To demonstrate ASER’s coverage on the knowledge in CSKBs, we present the proportion of heads, tails, and edges that can be found in the ASER$_{norm}$ via exact string match in Table 4.3. For edges, we report the proportion of edges where the corresponding heads and tails can be connected by a path in ASER. We also report the average shortest path length in ASER for those matched edges from the CSKB in the #hops column, showing that ASER can entail such commonsense knowledge within several hops of path reasoning, which builds the foundation of commonsense reasoning on

	ASER$_{norm}$				Avg. Degree in ASER$_{norm}$				Avg. Degree in $\mathcal{C}$			
	Coverage				In-Degree		Out-Degree		In-Degree		Out-Degree	
	head(%)	tail(%)	edge(%)	#hops	head	tail	head	tail	head	tail	head	tail
ATOMIC	79.76	77.11	59.32	2.57	90.9	61.3	91.2	61.6	4.2	3.4	34.6	1.5
ATOMIC$_{20}^{20}$	80.39	47.33	36.73	2.65	96.9	66.9	97.3	67.3	4.3	2.9	34.6	1.5
ConceptNet	77.72	54.79	43.51	2.37	210.7	88.9	211.6	88.9	15.1	8.0	26.2	4.1
GLUCOSE	91.48	91.85	81.01	2.37	224.9	246.4	226.6	248.0	7.2	7.7	6.7	5.5

Table 4.3: The overall matching statistics for the four CSKBs. The *edge* column indicates the proportion of edges where their heads and tails can be connected by paths in ASER. Average (in and out)-degree on ASER$_{norm}$ and $\mathcal{C}$ for nodes from the CSKBs is also presented. The statistics in $\mathcal{C}$ is different from [2] as we check the degree on the aligned CSKB $\mathcal{C}$ instead of each individual CSKB.

ASER. In addition, the average degree in $\mathcal{G}^c$ and $\mathcal{C}$ for heads and tails from each CSKB is also presented in the table. The total number of triples for each relation in the CSKBs is presented in Table 4.4. There are 18 commonsense relations in total for CSKBs and 15 relations in ASER. More detailed descriptions and examples of the unification are presented in the Appendix (Table A.1, A.2).

4.2.6 Evaluation Set Preparation

For the ground truth commonsense triples from the CSKBs, we split them into train, development, and test set with the proportion 8:1:1. Negative examples are sampled by selecting a random head and a random tail from the aligned $\mathcal{G}^c$ such that the ratio of negative and ground truth triples is 1:1. To form a diverse evaluation set, we sample 20K triples from the original automatically constructed test set (denoted as "*Original Test Set*"), 20K from the edges in ASER where heads come from CSKBs and tails are from ASER (denoted as "*CSKB head + ASER tail*"), and 20K triples in ASER where both heads and tails come from ASER (denoted as "*ASER edges*"). The distribution of different relations in this evaluation set is the same as in the original test set. The sampled evaluation set is then annotated to acquire ground labels.

4.3 Human Annotation

4.3.1 Setups

The human annotation is carried out on Amazon Mechanical Turk. Workers are provided with sentences in the form of natural language translated from knowledge triples (e.g., for `xReact`, an (h, r, t) triple is translated to "If h, then, PersonX feels t"). Additionally, following ATOMIC$_{20}^{20}$ [10], annotators are asked to rate each triple in a four-point Likert scale: *Always/Often*, *Sometimes/Likely*, *Farfetched/Never*, and *Invalid*. Triples receiving the former two labels will be treated as *Plausible* or otherwise *Implausible*. Each HIT (task) includes 10 triples with the same relation type, and each sentence is labeled by 5 workers. We take the majority vote among five votes as the final result for each triple. To

Relation	ATOMIC$\binom{20}{20}$	ConceptNet	GLUCOSE
oEffect	21,497	0	7,595
xEffect	61,021	0	30,596
gEffect	0	0	8,577
oWant	35,477	0	1,766
xWant	83,776	0	11,439
gWant	0	0	5,138
oReact	21,110	0	3,077
xReact	50,535	0	13,203
gReact	0	0	2,683
xAttr	89,337	0	7,664
xNeed	61,487	0	0
xIntent	29,034	0	8,292
isBefore	18,798	0	0
isAfter	18,600	0	0
HinderedBy	87,580	0	0
xReason	189	0	0
Causes	0	42	26,746
HasSubEvent	0	9,934	0
Total	578,252	10,165	126,776

Table 4.4: Relation distribution statistics for different CSKBs. Due to the filter in Section 4.2.1, the statistics are different from the original papers.

avoid ambiguity and control the quality, we finalize the dataset by selecting triples where workers reach an agreement on at least 4 votes.

4.3.2 Quality Control

For strict quality control, we carry out two rounds of qualification tests to select workers and provide a special training round. First, workers satisfying the following requirements are invited to participate in our qualification tests: 1) at least 1K HITs approved, and 2) at least 95% approval rate. Second, a qualification question set including both straightforward and tricky questions is created by experts, who are graduate students in HKUST and have a clear understanding of this task. 760 triples sampled from the original dataset are annotated by the experts. Each worker needs to answer a HIT containing 10 questions from the qualification set and their answers are compared with the expert annotation. Annotators who correctly answer at least 8 out of 10 questions are selected in the second round. 671 workers participated in the qualification test, among which 141 (21.01%) workers are selected as our main round annotators. To further enhance the quality, we carry out an extra training round

	Dev	Test	Train
# Triples	6,217	25,514	1,100,362
% Plausible	51.05%	51.74%	-
% Novel Nodes	67.40%	70.01%	-

Table 4.5: Statistics of the annotated evaluation set. # triples indicates the number of triples in the dataset, % Plausible indicates the proportion of plausible triples after majority voting, and % Novel Nodes is the proportion of nodes that do not appear in the training CSKBs. We also report the scale of the unannotated training set (including random negative examples) for reference.

for the main round annotators. For each relation, annotators are asked to rate 10 tricky triples carefully selected by experts. A grading report with detailed explanations on every triple is sent to all workers afterward to help them fully understand the annotation task.

After filtering, we acquire human-annotated labels for 31,731 triples. The IAA score is 71.51% calculated using pairwise agreement proportion, and the Fleiss's κ [159] is 0.43. We further split the proportion of the development set and test set as 2:8. The overall statistics of this evaluation set are presented in Table 4.5. To acquire human performance, we sample 5% of the triples from the test set, and ask experts as introduced above to provide two additional votes for the triples. The agreement between labels acquired by majority voting and the 5+2 annotation labels is used as the final human performance of this task.

4.4 Experiments

In this section, we introduce the baselines and our proposed model KG-BERTSAGE for the CSKB population task, as well as the experimental setups.

4.4.1 Model

The objective of a population model is to determine the plausibility of an (h, r, t) triple, where nodes can frequently be out of the domain of the training set. In this sense, transductive methods based on knowledge base embeddings [2] are not studied here. We present several ways of encoding triples in an inductive manner.

BERT. The embeddings of h, r, t are encoded as the embeddings of the [CLS] tokens after feeding them separately as sentences to BERT. For example, the relation `xReact` is encoded as the BERT embedding of "[CLS] xReact [SEP]". The embeddings are then concatenated as the final representation of the triple, $[s_h, s_r, s_t]$.

BERTSAGE. The idea of BERTSAGE [110] is to leverage the neighbor information of nodes through a graph neural network layer for their final embedding. For h, denote its BERT embedding as s_h, then the final embedding of h is $e_h = [s_h, \sum_{v \in \mathcal{N}(h)} s_v / |\mathcal{N}(h)|]$, where $\mathcal{N}(h)$ is the neighbor function that returns the neighbors of h from $\mathcal{G}$. The final representation of the triple is then $[e_h, s_r, e_t]$.

KG-BERT. KG-BERT(a) [111] encodes a triple by concatenating the elements in (h, r, t) into a single sentence and encode it with BERT. Specifically, the input is the string concatenation of [CLS], h, [SEP], r, [SEP], t, and [SEP].

KG-BERTSAGE. As KG-BERT doesn't take into account graph structures directly, we propose to add an additional graph SAmpling and AGgregation layer [116] to better learn the graph structures. Specifically, denoting the embedding of the (h, r, t) triple by KG-BERT as KG-BERT(h, r, t), the model of KG-BERTSAGE is the concatenation of KG-BERT(h, r, t), $\sum_{(r',v) \in \mathcal{N}(h)}$ KG-BERT $(h, r', v)/|\mathcal{N}(h)|$, and $\sum_{(r',v) \in \mathcal{N}(t)}$ KG-BERT $(v, r', t)/|\mathcal{N}(t)|$. Here, $\mathcal{N}(h)$ returns the neighboring edges of node h.

Relation	xWnt	oWnt	gWnt	xEfct	oEfct	gEfct	xRct	oRct	gRct	xAttr	xInt	xNeed	Cause	xRsn	isBfr	isAft	Hndr.	HasSubE.	all
BERT	57.7	64.9	66.3	59.1	66.2	60.0	50.6	**68.7**	72.3	56.2	å 63.9	56.4	48.3	34.5	59.2	58.0	66.1	73.0	59.4
BERTSAGE	54.7	58.9	58.0	58.0	70.0	54.7	52.8	62.4	**76.6**	55.0	61.0	57.1	46.2	45.5	66.7	64.9	69.6	**80.4**	60.0
KG-BERT	63.2	**69.8**	**69.0**	68.0	70.6	61.0	57.0	64.0	73.8	**59.5**	**64.9**	64.6	47.4	**90.9**	78.0	**77.5**	75.9	68.5	66.1
KG-BERTSAGE	**66.0**	68.9	68.6	**68.2**	**70.8**	**62.3**	**60.5**	64.6	74.1	59.1	63.0	**65.4**	**50.0**	76.4	**78.2**	77.4	**77.5**	67.0	**67.2**
Human	86.2	86.8	83.3	85.2	83.9	79.8	81.1	82.6	76.5	82.6	85.6	87.4	80.1	73.7	89.8	89.9	85.3	85.7	84.4

Table 4.6: Experimental results on CSKB population. We report the AUC (×100) here for each relation. The improvement under "all" is statistically significant using Randomization Test [3], with $p < 0.05$.

Relation	#Eval.	#Train
xWant	2,605	152,634
oWant	999	59,688
gWant	207	8,093
xEffect	2,757	144,799
oEffect	667	46,555
gEffect	287	13,529
xReact	2,999	100,853
oReact	921	38,581
gReact	164	4,169
xAttr	2,561	152,949
xIntent	1,017	59,138
xNeed	1,532	98,830
Causes	1,422	40,450
xReason	16	320
isBefore	879	27,784
isAfter	1,152	27,414
HinderedBy	4,870	127,320
HasSubEvent	459	16,410

Table 4.7: Number of triples of each relation in the Eval. (dev+test) and Train set.

4.4.2 Setup

We train the population model using a triple classification task, where ground truth triples come from the original CSKB, and the negative examples are randomly sampled from the aligned graph $\mathcal{G}^c$. The model needs to discriminate whether an (h, r, t) triple in the human-annotated evaluation set is plausible or not. For evaluation, we use the AUC score as the evaluation metric, as this commonsense reasoning task is essentially a ranking task that is expected to rank plausible assertions higher than those farfetched assertions.

We use BERT$_{base}$ from the Transformer[1] library, and use learning rate 5×10^{-5} and batch size 32 for all models. The statistics of each relation are shown in Table 4.7. We select the best models individually for each relation based on the corresponding development set. Besides AUC scores for each relation, we also report the AUC score for all relations by the weighted sum of the break-down scores, weighted by the proportion of test examples of the relation. This is reasonable as AUC essentially represents the probability that a positive example will be ranked higher than a negative example.

[1] https://transformer.huggingface.co/

4.4.3 Main Results

The main experimental results are shown in Table 4.6. KG-BERTSAGE performs the best among all, as it both encodes an (h, r, t) as a whole and takes full advantage of neighboring information in the graph. Moreover, all models are significantly lower than human performance with a relatively large margin.

ASER can on the one hand provide candidate triples for populating CSKBs, and can on the other hand provide graph structure for learning commonsense reasoning. From the average degree in Table 4.3, the graph acquired by grounding CSKBs to ASER can provide far more neighbor information than using the CSKBs only. While KG-BERT treats the task directly as a simple triple classification task and takes only the triples as input, it does not explicitly take into consideration the graph structure. KG-BERTSAGE on the other hand leverages an additional GraphSAGE layer to aggregate the graph information from ASER, thus achieving better performance. It demonstrates that it is beneficial to incorporate those un-annotated ASER graph structures where multiple-hop paths are grounded between commonsense heads and tails. Though BERTSAGE also incorporates neighboring information, it only leverages the ASER nodes representation and ignores the complete relational information of triples as KG-BERTSAGE does. As a result, it doesn't outperform BERT by much for the task.

Relation	xWnt	oWnt	gWnt	xEfct	oEfct	gEfct	xRct	oRct	gRct	xAttr	xInt	xNeed	Cause	xRsn	isBfr	isAft	Hndr.	HasSubE.	all
KG-BERT																			
· on ATOMIC$\binom{20}{20}$	61.0	64.2	68.0	62.9	67.1	**64.8**	**58.8**	60.2	68.6	58.9	62.4	63.7	**55.8**	58.2	77.7	76.7	75.5	67.6	65.2
· on GLUCOSE	62.3	67.6	**69.2**	61.6	**71.5**	57.3	58.0	63.4	**77.0**	57.7	61.0	50.4	48.1	72.7	61.0	50.6	59.2	68.0	59.2
· on ConceptNet	58.0	62.0	59.4	56.2	52.5	61.4	52.3	57.0	54.4	57.1	61.8	57.4	55.6	78.2	61.8	60.8	63.2	60.9	58.3
· on all	**63.2**	**69.8**	69.0	**68.0**	70.6	61.0	57.0	**64.0**	73.8	**59.5**	**64.9**	**64.6**	47.4	**90.9**	**78.0**	**77.5**	**75.9**	**68.5**	**66.1**
KG-BERTSAGE																			
· on ATOMIC$\binom{20}{20}$	63.1	64.7	65.6	63.7	67.5	**65.7**	56.1	60.3	64.9	56.8	60.5	63.7	**56.5**	65.5	76.9	76.6	76.9	63.8	65.1
· on GLUCOSE	61.7	68.3	**70.8**	61.1	**71.9**	60.1	56.1	61.4	71.3	56.5	60.5	46.8	50.5	69.1	60.6	51.7	60.0	**72.4**	58.9
· on ConceptNet	57.7	55.0	59.8	60.1	57.3	62.2	50.2	50.9	50.9	52.3	56.8	52.1	52.6	70.9	53.8	44.5	58.3	59.8	55.0
· on all	**66.0**	**68.9**	68.6	**68.2**	70.8	62.3	**60.5**	**64.6**	**74.1**	**59.1**	**63.0**	**65.4**	50.0	**76.4**	**78.2**	**77.4**	**77.5**	67.0	**67.2**

Table 4.8: Effects of different training sets.

Model	*Original Test Set*	*CSKB head + ASER tail*	*ASER edges*
BERT	65.0	47.9	44.6
BERTSAGE	67.2	49.4	46.2
KG-BERT	77.8	55.2	50.3
KG-BERTSAGE	**78.2**	**57.5**	**52.3**

Table 4.9: AUC scores grouped by the types of the evaluation sets defined in 4.2.6. The latter two groups are harder for neural models to distinguish.

4.4.4 Zero-shot Setting

We also investigate the effects of different training CSKBs as shown in Table 4.8. Models are then trained on the graphs only consisting of commonsense knowledge from ATOMIC($^{20}_{20}$), GLUCOSE, and ConceptNet, respectively. The models trained on all CSKBs achieve better performance both for each individual relation and on the whole. We can conclude that more high-quality commonsense triples for training from diverse dimensions can benefit the performance of such commonsense reasoning.

When trained on each CSKB dataset, there are some relations that are never seen in the training set. As all of the models use BERT to encode relations, the models are *inductive* and can thus reason triples for unseen relations in a zero-shot setting. For example, the `isBefore` and `isAfter` relations are not presented in GLUCOSE, while after training KG-BERTSAGE on GLUCOSE, it can still achieve fair AUC scores. Though not trained explicitly on the `isBefore` and `isAfter` relations, the model can transfer the knowledge from other relations and apply them to the unseen ones.

4.5 Error Analysis

As defined in Section 4.2.6, the evaluation set is composed of three parts: edges coming from the original test set (*Original Test Set*), edges where heads come from CSKBs and tails from ASER (*CSKB head + ASER tail*), and edges from the whole ASER graph (*ASER edges*). The break-down AUC scores of different groups given all models are shown in Table 4.9. The performances under the *Original Test Set* of all models are remarkably better than the other two groups, as the edges in the original test set are from the same

Head	Relation	Tail	Label	Pred.
PersonX go to nurse	`xEffect`	*PersonX* use to get headache	0	1
PersonX have a quiz	`Causes`	*PersonX* have pen	0	1
PersonX be strong	`oWant`	*PersonY* like *PersonX*	0	1
PersonX feel a pain	`xIntent`	PersonX finger have be chop off	0	1

Table 4.10: Examples of error predictions made by KG-BERTSAGE, where the head and tail are semantically related while not conformed to the designated commonsense relation.

domain as the training examples. The other two groups, where there are more unseen nodes and edges, are harder for the neural models to distinguish. The results show that simple commonsense reasoning models struggle to be generalized to unseen nodes and edges. As a result, in order to improve the performance of this CSKB population task, more attention should be paid to the generalization ability of commonsense reasoning on unseen nodes and edges.

Moreover, by taking a brief inspection of the test set, we found that errors occur when encountering triples that are not logically sound but semantically related. Some examples are presented in Table 4.10. For the triple (*PersonX* go to nurse, `xEffect`, *PersonX* use to get headache), the head event and tail event are highly related. However, the fact that someone gets a headache should be the reason instead of the result of going to the nurse. More similar errors are presented in the rest of the table. These failures may be because when using BERT-based models, the training may not be well performed for the logical relations or discourse but still recognize the semantic relatedness patterns.

4.6 CKBP v2

However, there is concern regarding the quality of the above crowdsourced CKBP dataset, denoted as CKBP v1. CKBP v1 instances are randomly sampled from the whole population space, resulting in a low recall of plausible commonsense knowledge due to the noise in candidate discourse knowledge. Moreover, as pointed out by [160], current crowdsourced commonsense benchmarks often contain a substantial fraction of incorrect answers; we also find it true for CKBP v1 after manual inspection. For example, annotators frequently make mistakes on some subtle relations such as `xIntent`, which should describe an *intention* instead of a *consequence*.

Therefore, to address the quality issue, we present a more high-quality and adversarially constructed evaluation set by expert annotation. Leveraging the existing framework, we build CKBP v2 by randomly sampling 2.5k instances from CKBP v1 and adding 2.5k adversarial instances, leading to a total of 5k instances as an evaluation set. These instances are then annotated by experts with substantial expertise in machine commonsense. Then, we present both intrinsic and extrinsic experiments based on CKBP v2. We study the performance of both supervised and semi-supervised task-specific models, together with powerful off-the-shelf language models, such as ChatGPT [41] and Vera [161], and show that the CKBP v2 evaluation set is still challenging even for advanced language models. Moreover, by employing a CSKB Population model that demonstrates satisfactory performance on CKBP v2, we can enrich existing CSKBs with diverse and novel knowledge that significantly benefits downstream reasoning. We present methodologies and experiments on generative commonsense inference [56] and zero-shot commonsense question answering [162], and show that the acquired commonsense knowledge can be valuable augmented data on the original CSKB and lead to improved downstream performance. In particular, CKBP v2-preferred population model exhibits better alignment than CKBP v1 with advancements in generative commonsense inference.

4.6.1 Dataset Preparation

We randomly sampled 2.5k instances from CKBP v1 and 2.5k adversarial instances to form CKBP v2. Instances from CKBP v1 are sampled so that the ratio of the number of triples between relations remains unchanged. Meanwhile, the adversarial instances are ones from the candidate knowledge base ASER that the finetuned baseline KG-BERT [117] model confidently believes they are plausible, i.e., receives plausibility score ≥ 0.9. To ensure the diversity of adversarial instances and hence the evaluation set, we adopt an additional diversity filter using self-BLEU following [102]. The triples annotated as negative are considered *hard negatives* as they are what a standard CSKB Population model would favor. Note that we only consider instances of 15 relations other than `general Want/React/Effect`,

	# Triples	% Plau.	% Unseen
split			
Dev	958	20.46	56.79
Test	4,048	22.06	60.43
instance type			
In-Domain	845	34.56	43.79
Out-of-Domain	1,653	11.92	63.37
Adv.	2,508	23.92	61.12
relation			
xWant	611	22.75	54.01
oWant	239	25.94	58.18
xEffect	603	29.68	55.23
oEffect	172	21.51	58.91
xReact	533	20.64	51.18
oReact	183	13.66	50.70
xAttr	605	23.47	52.91
xIntent	239	16.32	58.40
xNeed	378	25.66	55.37
Causes	236	21.61	55.41
xReason	5	40.0	30.0
isBefore	157	28.03	54.80
isAfter	182	24.73	55.40
HinderedBy	777	12.1	63.17
HasSubEvent	86	26.74	61.04

Table 4.11: Statistics of CKBP v2. # Triples, $\%$ Plausible, and $\%$ Unseen, respectively, indicate the number of triples in the subset, the proportion of plausible triples after label finalization, and the proportion of nodes that do not appear in the training set.

because most of the triples on the three relations are broken sentences in CKBP v1. We also removed samples of these relations from the training set.

4.6.2 Annotation Process

Setup We recruited four human experts for the annotation work. The experts are graduate NLP researchers with at least one year of experience working on CSKBs. We randomly divide 5k samples into 4 parts, then for i from 0 to 3, assign the i^{th} and $(i+1 \bmod 4)^{th}$ parts to the i^{th} expert. In this way, two different annotators annotate each triple, and we can fully compare the pairwise agreement between all four annotators. Experts are provided with knowledge triples in the format of (h, r, t), referencing the definition and examples of all relations in ATOMIC$^{20}_{20}$ [10]. We ask annotators to judge the plausibility of triples in a three-point Likert scale with corresponding scores: Always/Often (1), Sometimes (0.5), Rarely/Never/Ambiguous/Invalid (0). The final label of an instance is determined as *plausible* if and only if it receives at least one score of 1 and the other score is at least 0.5. For remaining cases, the final label is *implausible*. After finalizing the annotation, we split the evaluation set into development and test sets with a ratio of 1:4 with the preservation of distribution w.r.t labels, relations, and instance types. To estimate human performance, we treat expert annotations as two sets of predictions and compare them to the final labels.

Similar to CKBP v1, we categorize the evaluation set into three groups based on their origin, which are 1) ID: in-domain, whose head and tail events are all from CSKBs, 2) OOD: out-of-domain, which has at least one event outside of CSKBs (equivalent to "CSKB head + ASER tail" and "ASER Edges" in CKBP v1), and 3) *Adv.*: adversarial examples newly introduced in CKBP v2.

Quality Control Although annotators are experts with a clear understanding of the CSKB Population, we acknowledge the ambiguity of CSKB relations and the difficulty in discriminating between them. To control the quality, we provide guidance as a list of scoring criteria. We also carried out a dry run, which asked them to annotate 60 instances covering all relations in order to establish a unified understanding of the problem among participants.

After that, we carry out the main round, where the annotators perform their jobs individually and independently. Throughout the process, we regularly conduct random checks on the samples and engage in discussions with annotators to address any disagreements. We then use the insights gained from these discussions to update and refine our guidance iteratively. After the individual annotation, we facilitated a conflict resolution session to address instances with contrasting scores of 1 and 0. After resolving conflicts, we have the average inter-annotator agreement score IAA as 90.55%.

Category	Model	AUC				F1			
		all	ID	OOD	*Adv.*	all	ID	OOD	*Adv.*
Zero-shot	GPT2-large	56.47	56.60	58.31	54.22	35.37	47.40	24.06	36.84
	GPT2-XL	56.79	54.47	56.70	54.63	35.22	47.62	23.49	36.65
	GPT3 `text-davinci-003`	61.63	65.93	59.17	59.98	39.44	51.09	28.57	38.20
	ChatGPT `gpt-3.5-turbo`	65.77	70.37	62.56	62.27	45.93	62.59	44.79	26.86
Supervised Learning	KG-BERT (BERT-base)	71.33	84.60	64.47	62.9	45.03	69.27	26.53	41.97
	KG-BERT (RoBERTa-L)	73.70	85.53	67.70	65.60	46.70	69.73	30.73	43.27
	COMET (GPT2-L)	70.00	79.02	66.43	62.62	45.55	61.90	32.14	42.15
	COMET (GPT2-XL)	70.32	79.66	66.53	63.22	45.32	63.34	31.18	40.83
	Vera (T5-xxlarge)	72.45	78.84	68.40	**68.16**	**52.13**	**71.73**	**36.74**	**50.02**
Semi-Supervised	PseudoReasoner BERT-base	71.93	84.23	66.67	63.43	45.47	68.67	30.17	41.77
	PseudoReasoner RoBERTa-L	**74.33**	**85.57**	**69.33**	66.37	46.63	69.70	30.87	43.13
Human		94.1	94.9	91.4	94.5	91.5	94.3	86.9	91.5

Table 4.12: Main experimental results on CKBP v2. Both AUC and F1 are used as evaluation metrics. The "all" column indicates the overall performance, and ID, OOD, *Adv.* indicate the performance of the In-domain, Out-of-domain, and Adversarial subset. The best results are **boldfaced**, and the second-best ones are underlined.

4.6.3 Data Analysis

The overall statistics of CKBP v2 are shown in Table 4.11. It can be easily observed that the new evaluation set has data imbalance issues. However, we do not down-sample the evaluation set to achieve the data balance since the imbalance better reflects the true distribution of plausible and implausible commonsense knowledge in ASER. Given this imbalance, we notice that the AUC scores of examined population models will naturally be high. Also, in the real application of population models, we focus on the precision and recall of the detection for plausible commonsense instances. Thus, in Section 4.6.4, along with AUC, we also report the binary F1 scores for each experimented model.

4.6.4 Experiments

Setup We examine several models which were previously evaluated on CKBP v1, including zero-shot GPT models [37], supervised-learning baselines KG-BERT [117] and COMET [56], and semi-supervised-learning models PseudoReasoner [118] with two backbone encoders, BERT-base-uncased [35] and RoBERTa-large [163]. We use Huggingface[2] Transformers [164] to build our code base. For discriminative models, we set the learning rate as 1e-5, batch size 64/32 for base/large variants, respectively, and the number of training epochs as 1. For generative models (COMET), we use learning rate 1e-5 and batch size 32 to train in 3 epochs. Negative perplexity scores are used as the final prediction scores. For PseudoReasoner, we adopt the best settings in [118], where we first finetune the KG-BERT model on pseudo-labeling data for one epoch, then from the best checkpoint, we resume the finetuning process on the original training data. Note that the training data and unlabeled data are taken from PseudoReasoner [118]. We run each baseline three times with different random seeds, then average the result and report in Table 4.12. For GPT3 [100] and ChatGPT experiments, we use simple prompts asking them to decide whether an assertion is plausible or not.

Result and Analysis The results are shown in Table 4.12. We provide the AUC score and F1 score of all the baselines on the test set in terms of overall performance (all), performance on the subset of ID, OOD, and *Adv.* samples. When calculating F1, for discriminative models, we set the decision threshold as 0.5 (as default), while for generative models, as perplexity serves as the final prediction score, we tune the threshold to obtain the highest F1 score on the development set for each run.

In the zero-shot setting, the scores increase by the version of GPT. GPT3 [3] gives a significant improvement over GPT2 models, and ChatGPT surpasses its sibling GPT3 with a similar margin of improvement. Nonetheless, despite the performance improvement from ChatGPT, there is still a clear gap between the zero-shot and (semi-)supervised settings.

In terms of supervised and semi-supervised learning, we observe different scenarios between KG-BERT's performance and COMET's performance, comparing to the result on CKBP v1 reported in PseudoReasoner [118]. Here, on CKBP v2, KG-BERT outperforms COMET with a significant gap of 3 AUC overall and also outperforms in all subsets of the test set. This shows the importance of including negative (implausible) examples in the

[2] https://huggingface.co/

[3] `text-davinci-003`

training for discriminating commonsense. This also explains why there is no significant improvement of PseudoReasoner over the baseline KG-BERT on this new evaluation set.

4.7 Conclusion

In this section, we benchmark the CSKB population task by proposing a dataset by aligning four popular CSKBs and an eventuality graph ASER, and provide a high-quality human-annotated evaluation set to test models' reasoning ability. We also propose KG-BERTSAGE to both incorporate the semantic of knowledge triples and the subgraph structure to conduct reasoning, which achieves the best performance among other counterparts. Experimental results also show that the task of reasoning unseen triples outside of the domain of CSKB is a hard task where current models are far away from human performance, which brings challenges to the community for future research.

CHAPTER 5

SEMI-SUPERVISED LEARNING FOR IMPROVED COMMONSENSE KNOWLEDGE BASES POPULATION

5.1 Preliminary

Commonsense knowledge are the common agreements by most people on daily entities, which are crucial for intelligent systems to act sensibly in the real world [14, 12]. Endowing natural language understanding systems with the ability to draw commonsense reasoning remains an important yet challenging task.

Throughout the development of automated commonsense understanding, CommonSense Knowledge Base (CSKB) has been an important form of automatic commonsense reasoning system that stores knowledge sources for drawing inferences. With expert-curated relations and human annotations, CSKBs such as ConceptNet [12], ATOMIC [55, 10], and GLUCOSE [1] are developed to study commonsense regarding properties of objects, causes and effects of events and activities, motivations and emotional trajectories of humans on certain circumstances, and so on. As those human-annotated CSKBs are sparse and usually of a small scale and coverage, reasoning tasks on CSKB such as *CSKB Completion* [165, 166, 167] and *CSKB Population* [110, 168] are defined with the goal of either adding new edges/assertions within the training knowledge base (*CSKB Completion*), or adding new edges/assertions from outside of CSKBs (*CSKB Population*). A visualized comparison between the two tasks is shown in Figure 5.1.

Different from CSKB Completion, which adopts a *close-world assumption* and assumes all knowledge is in-domain, the population task deals with unseen entities and requires a more out-of-distribution reasoning ability. In this section, we study commonsense reasoning in the context of CSKB Population. In this task, four mainstream CSKBs,

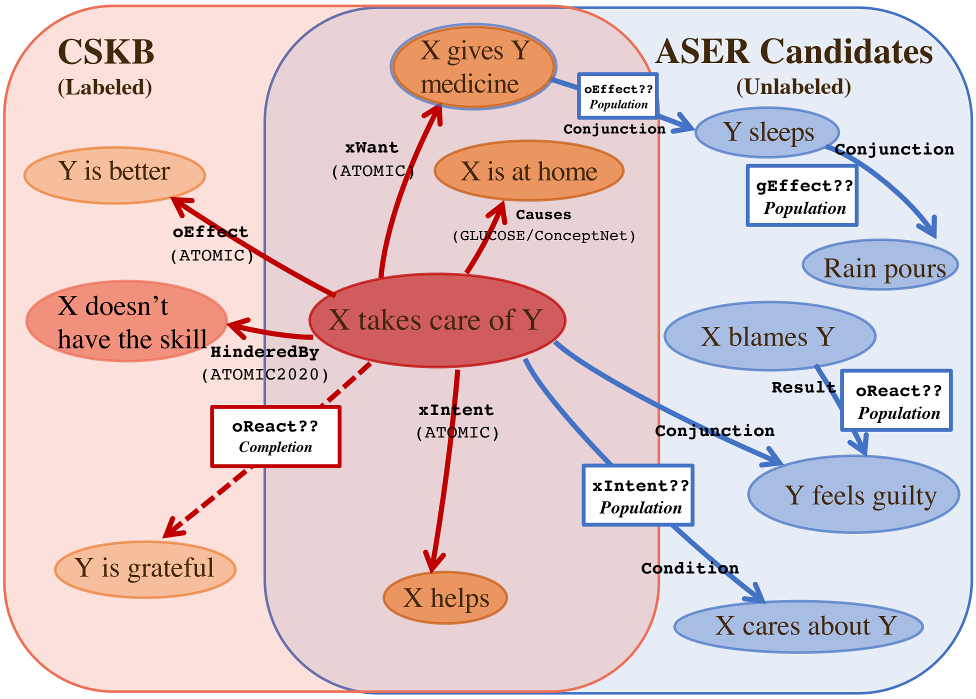


Figure 5.1: An example of CSKB Population. The coral part (left) and the blue part (right) respectively represent the labeled CSKBs and the unlabeled candidate pool. The entities in the overlap parts are marked with coral shape and blue outline. The reasoning within the CSKB (coral-outlined boxes) belongs to the CSKB Completion part, and the reasoning that are not limited within the domain of CSKBs belongs to CSKB Population.

ConceptNet, ATOMIC, ATOMIC$^{20}_{20}$, and GLUCOSE are aligned together as the labeled dataset. ASER [59], a large-scale eventuality (events, activities, and states) knowledge graph is aligned with CSKBs and serves as unlabeled candidates for populating commonsense knowledge. Human annotations on held-out dev/test sets sampled from both CSKBs and ASER are provided as the evaluation set.

There are two major challenges remaining unsolved for CSKB Population. First, the scale of the annotated training set (ConceptNet, ATOMIC, and GLUCOSE) is approximately 1M samples, too small compared with 200M of the actual candidate space to perform population (ASER). Second, as inherently only ground-truth (positive) examples are provided by CSKBs, the randomly sampled negative examples in the task are less informative and may lead the model to overfit artifacts of the dataset. A supervised learning model finetuned on such an annotated training set is hard to generalize to out-of-domain knowledge space, as shown in the experiments in Section 4 and also Table 5.3, where the AUC for out-of-domain test sets performs over 10 points worse than the in-domain part.

To address the above challenges, we propose **PseudoReasoner**, a semi-supervised learning framework that uses a pre-trained commonsense teacher model to automatically label the unlabeled candidates to serve as pseudo labels, such that the student model can be further finetuned with pseudo labels to improve out-of-domain commonsense reasoning ability. In fact, pre-trained language models finetuned on commonsense knowledge bases have been shown to perform generalizable commonsense reasoning on downstream tasks to some extent. For example, leveraging commonsense knowledge generated by COMET [56], a language model finetuned on ATOMIC, can improve the performance on commonsense QA [6, 169, 170]. Different from the text generation paradigm as in previous works, here we leverage the commonsense language model as a teacher model for labeling unlabeled candidates. To further improve the quality of pseudo labels, we use both influence function [171] and the student model's prediction to select highly confident pseudo examples.

Our contribution is three-fold:

1. We introduce a new way of providing pseudo labels for CSKB Population by leveraging generative commonsense language models.
2. We propose a semi-supervised learning framework with pseudo labels and a special filtering mechanism based on influence function and student model's prediction that significantly improve the performance of CSKB Population, especially for out-of-domain knowledge triples.
3. We demonstrate the effectiveness of our framework by extensive experiments on different backbone models and different semi-supervised learning methods. We achieve the state-of-the-art performance on this task.

5.1.1 Problem Definition

Denote a labeled ground-truth Commonsense Knowledge Base as $D_l^+ = \{(h, r, t) | h \in H_C, r \in R, t \in T_C\}$. The overall labeled dataset $D_l = \{(h, r, t), y)\}$ is composed of D_l^+ and a randomly sampled negative dataset D_l^- from the CSKB, where $y \in \{0, 1\}$ is the label of the triple. Triples from D_l^+ are labeled 1 while those from D_l^- are labeled 0. H_C and T_C are the set of heads and tails in the CSKB. R is the relation set.

	Original Test Set	*CSKB head + ASER tail*	*ASER edges*
Sampled from	D_l	D_l head, D_u tail	D_u
Domain	In-domain	Out-of-domain	Out-of-domain
# triples (dev)	2,042	2,193	1,982
# triples (test)	8,437	9,103	7,974

Table 5.1: The details of evaluation set categorization.

	D_l	D_u
# triples	1,119,517 (training)	218,809,746

Table 5.2: Statistics of labeled D_l (CSKB with negative examples) and unlabeled D_u (processed ASER).

$D_u = \{(h, r, t)|h \in H_u, r \in R, t \in T_u\}$ is an unlabeled candidate knowledge base of the same format and relation set as the CSKB. It is of a much larger scale than the labeled part and is a source for populating commonsense knowledge. H_u and T_u are the set of heads and tails in the unlabeled KB. The task is defined as given the labeled commonsense knowledge base D_l as training source, predict the plausibility of triples from D_u.

We use the dataset provided by DISCOS [115]. Here, the set of heads H_C, tails T_C, and relations R come from the alignment of ConceptNet, ATOMIC, ATOMIC$^{20}_{20}$ (newly developed relations), and GLUCOSE, as in the original paper. The unlabeled KB D_u is adapted from ASER, where the discourse relations are converted to commonsense relations to serve as candidates for population. The evaluation dataset with 32K triples is sampled from both D_l and D_u and manually annotated. There are three categories of the evaluation set, *Original Test Set* , *CSKB head + ASER tail*, and *ASER edges*, where the first category is sampled from the held-out test split in D_l (both D_l^+ and negative examples D_l^-) and is thus an in-domain test set, and the latter two are novel assertions outside of D_l and are thus out-of-domain. The statistics and descriptions of the training and evaluation datasets are shown in Table 5.2 and 5.1.

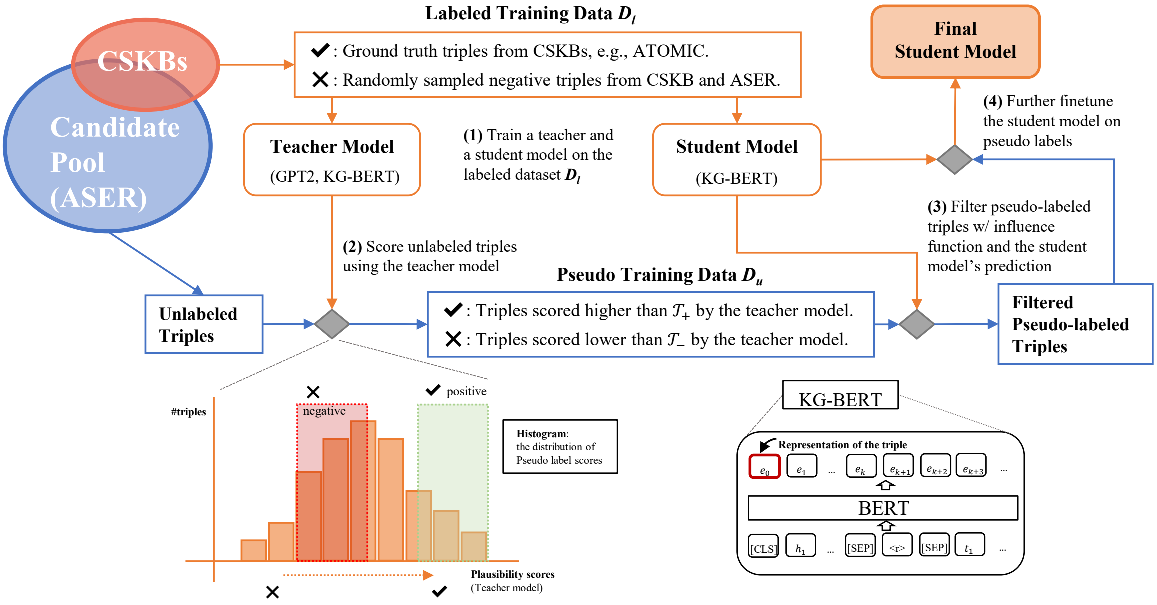

CSKBs
Candidate Pool (ASER)
Labeled Training Data D_l
✓: Ground truth triples from CSKBs, e.g., ATOMIC.
✗: Randomly sampled negative triples from CSKB and ASER.
Final Student Model
Teacher Model (GPT2, KG-BERT)
(1) Train a teacher and a student model on the labeled dataset D_l
Student Model (KG-BERT)
(4) Further finetune the student model on pseudo labels
(3) Filter pseudo-labeled triples w/ influence function and the student model's prediction
(2) Score unlabeled triples using the teacher model
Pseudo Training Data D_u
Unlabeled Triples
✓: Triples scored higher than $\mathcal{T}_+$ by the teacher model.
✗: Triples scored lower than $\mathcal{T}_-$ by the teacher model.
Filtered Pseudo-labeled Triples
#triples
✗ negative
✓ positive
Histogram: the distribution of Pseudo label scores
Plausibility scores (Teacher model)
✗
✓
KG-BERT
Representation of the triple
e_0 e_1 ... e_k e_{k+1} e_{k+2} e_{k+3} ...
BERT
[CLS] h_1 ... [SEP] <r> [SEP] t_1 ...

Figure 5.2: An end-to-end workflow of **PseudoReasoner**. Four steps in the figure are elaborated in Section 5.3.

5.2 CSKB Population

5.2.1 Backbone Models

Considering the nature of the CSKB Population task is triple classification in the form of natural language, we use KG-BERT [4] as the backbone model. In detail, a triple (h, r, t) is concatenated and serialized as "[CLS], $h_1, ..., h_{|h|}$, [SEP], $[r]$, [SEP], $t_1, ..., t_{|t|}$". Here, [CLS] and [SEP] are the special tokens in BERT-based models [35]. [CLS] is used to represent the whole sentence, and [SEP] is used to separate different sentences respectively. $h_1, ..., h_{|h|}$ are the tokens of the head h, and $t_1, ..., t_{|t|}$ are the tokenized tokens of the tail t. $[r]$ is registered as a new special token for a certain relation r. After feeding the serialized version of (h, r, t) into a BERT-based masked language model, the representation of the special token [CLS] is regarded as the representation of the whole triple. It is trained to distinguish positive triples from negative triples with cross-entropy loss. Here x denotes a triple (h, r, t), P_L models the distribution of the labeled dataset D_l, and θ is the set of parameters for KG-BERT. $P_\theta(y|x)$ denotes the probability after feeding the model prediction logits to softmax under parameter set θ. Then the optimization objective is as follows:

$$J(\theta) = \mathbb{E}_{x_l \sim P_L(x)}[-\log P_\theta(y|x_l)]. \tag{5.1}$$

5.3 Methods

In this section, we present the details of the framework of PseudoReasoner. A sketch illustration of the model is presented in Figure 5.2. To sum up, the procedure of PseudoReasoner can be summarized into the following steps:

1. Train a teacher model and a student model on the labeled dataset D_l (Section 5.3.1).
2. Use the teacher model to predict plausibility scores on triples from the unlabeled D_u. Triples with high/low plausibility scores within pre-defined intervals are given label $1/0$ (Section 5.3.1).
3. Filter the pseudo labels with influence function with respect to the student model, and the student model's predictions. (Section 5.3.1).
4. Finetune the student model on filtered pseudo labels from 3). (Section 5.3.2).

5.3.1 Pseudo Label Construction

Teacher Models

We use a pre-trained teacher model on the labeled dataset to label the unlabeled triples. We define plausibility scores of an unlabeled triple x as $\alpha(x)$, where the higher the score the more plausible the triple is regarded by the teacher model. We choose two different forms of teacher models as follows:

• **GPT-2** [172]: As negative sampling in the labeled dataset D_l is noisy, we aim to use an alternative model that avoids the negative part D_l^-. We finetune a (COMET) GPT2 language model, as the representative of generative family, on the positive part of the labeled dataset, D_l^+, with a text generation task. For an (h, r, t) triple from D_l^+, denote x as the serialized version of the triple, "$h_1, ..., h_{|h|}, [r], t_1, ..., t_{|t|}$". θ_{LM} denotes the trainable parameters in GPT2 language model (LM). We minimize the negative log likelihood of each triple as indicated in Equation (5.2):

$$L(x, \theta_{LM}) = -\frac{1}{|x|}\sum_{i=1}^{|x|} \log P(x_i | x_{<i}, \theta_{LM}). \tag{5.2}$$

			In-domain	Out-of-domain (OOD)			
Category	Model	all	*Original Test Set*	*CSKB head + ASER tail*	*ASER edges*	Δ_{all}	Δ_{OOD}
Supervised Learning	KG-BERT (BERT-base) *110M*	62.5	74.2	51.9	54.7	-	-
	KG-BERT (BERT-large) *340M*	67.7	74.5	58.7	62.0	-	-
	KG-BERT (DeBERTa-base) *100M*	64.5	73.2	54.0	57.0	-	-
	KG-BERT (DeBERTa-large) *350M*	69.2	77.6	59.9	61.8	-	-
	KG-BERT (BART-base) *139M*	65.1	74.7	54.7	56.6	-	-
	KG-BERT (BART-large) *406M*	70.4	78.6	62.8	64.2	-	-
	KG-BERT (RoBERTa-base) *110M*	68.0	76.3	59.3	59.8	-	-
	KG-BERT (RoBERTa-large) *340M*	70.9	78.0	63.4	64.6	-	-
	COMET (GPT2-small) *117M*	69.6	71.6	67.4	65.0	-	-
	COMET (GPT2-medium) *345M*	69.7	71.9	67.0	67.9	-	-
	COMET (GPT2-large) *774M*	70.6	73.7	66.8	68.0	-	-
	COMET (GPT2-XL) *1558M*	70.7	74.6	66.7	67.6	-	-
Semi-supervised Learning (RoBERTa-large)	UDA (TF-IDF)	71.7	78.0	65.1	65.9	+0.8	+1.5
	UDA (back-trans.)	71.6	78.6	64.2	66.2	+0.7	+1.2
	G-DAUG	71.7	78.5	64.8	65.5	+0.8	+1.2
	G-DAUG (COMET-distill)	72.2	78.6	65.9	66.9	+1.3	+2.4
	Noisy-student	72.4	79.3	65.3	66.7	+1.5	+2.0
Ours	**PseudoReasoner** (BERT-base)	67.9	76.0	56.1	64.2	+5.4	+6.9
	PseudoReasoner (RoBERTa-large)	**74.2**	**80.1**	**69.5**	**69.3**	+3.3	+5.3

Table 5.3: Results on the test set of the CSKB Population benchmark. For supervised learning baselines, we report the result of KG-BERT [4] with four backbone encoders and GPT2 (use LM loss to score triples). For semi-supervised learning (SSL) baselines, we study UDA [5], G-DAUG [6], and Noisy-student [7]. The backbone encoders for SSL baselines are RoBERTa-large, which performs the best in the supervised setting. The number of parameters of backbone language models are presented as subscripts behind model names. Δ_{all} and Δ_{OOD} are the improvement on the "all" AUC and the Out-of-domain (OOD) AUC.

Denote the optimized parameters as θ^*_{LM}. Here, the plausibility function $\alpha(x) = -L(x, \theta^*_{LM})$, where the lower the loss, the higher the plausibility score by GPT2. Hence, for the triples from the unlabeled dataset, $\mathcal{D}_u$, we score every triple with Equation (5.2) on θ^*_{LM}.

• **KG-BERT**: Besides GPT2, KG-BERT itself, a discriminative model, can be used as a teacher model. This teacher model learns θ^* from the labeled dataset D_l with cross entropy loss in Equation (5.1). For an instance $\{(h, r, t), y\} \in D_l$, denote $x = (h, r, t)$, we use $\alpha(x)$ $= P_{\theta^*}(y{=}1|x)$ as x's plausibility score.

Acquiring Pseudo Labels

The triples whose plausibility scores $\alpha(x)$ are between $[\mathcal{T}^-_{min}, \mathcal{T}^-_{max}]$ are labeled as negative, and the triples within $[\mathcal{T}^+_{min}, \mathcal{T}^+_{max}]$ are labeled as positive. Here $\mathcal{T}^-_{min} < \mathcal{T}^-_{max} < \mathcal{T}^+_{min} < \mathcal{T}^+_{max}$. The reason that we introduce additional $\mathcal{T}^-_{min}$ and $\mathcal{T}^+_{max}$ is that we want to filter out the triples that are treated over plausible or implausible by GPT2 to reduce potential *selection bias*. For example, GPT2 has been shown to provide low loss for repetitive patterns instead of the plausibility of the semantics [100].

Pseudo Label Filters

To further improve the quality of pseudo labels, we propose two filtering mechanisms on pseudo labels for better finetuning.

• **Influence Function.** Filtering out detrimental training examples with influence function [171] can boost the model performance, as shown in [6] and [173]. A training example $z = ((h, r, t), y)$ will hurt the generalization ability of the model if including z in the training set results in a higher validation loss. Denote $L(\mathcal{Z}, \theta)$ as the loss function of dataset $\mathcal{Z}$ under the parameter set θ. Then the loss under training set $\mathcal{Z}_{train}$ is indicated in Equation (5.3):

$$L(\mathcal{Z}_{train}, \theta) = \frac{1}{|\mathcal{Z}_{train}|} \sum_{i=1}^{|\mathcal{Z}_{train}|} L(z_i, \theta). \tag{5.3}$$

Denote θ^* as the optimized parameters after training the model on $\mathcal{Z}_{train}$, and θ^*_{-z} as the optimized parameters after training the model on $\mathcal{Z}_{train} - \{z\}$. Denote $\mathcal{Z}_{val}$ as the validation set. The empirical criterion to determine z as a detrimental training example is Equation (5.4):

$$L(\mathcal{Z}_{val}, \theta^*) - L(\mathcal{Z}_{val}, \theta^*_{-z}) > 0. \tag{5.4}$$

The left-hand-side $L(\mathcal{Z}_{val}, \theta^*) - L(\mathcal{Z}_{val}, \theta^*_{-z})$ can be approximated without retraining the model by influence function [171]:

$$\mathcal{I}_{up,loss}(z) = -\nabla_\theta L(z_{val}, \theta^*)^\top H_{\theta^*}^{-1} \nabla_\theta L(z, \theta^*), \tag{5.5}$$

where $H_{\theta^*} = \frac{1}{|\mathcal{Z}_{train}|} \sum_{z_i \in \mathcal{Z}_{train}} \nabla^2_{\theta^*} L(z_i, \theta^*)$ is the Hessian. We linearly approximate $\mathcal{I}_{up,loss}$ with inverse hessian-vector product (HVP) introduced in LiSSA [174] following [171]. We filter out those examples with negative influence scores, which are harmful to the generalization of the model.

• **KG-BERT.** As the student model we use is KG-BERT, when the pseudo labels from

GPT2 are used, we can use the $P_{\theta^*}(y|x)$ produced from optimized KG-BERT as an additional filter to select pseudo labels. Specifically, pseudo labels $\{x = (h, r, t), y\}$ with $P_{\theta^*}(y|x) > 0.5$ are selected. This procedure can be viewed as ensembling GPT2 and KG-BERT.

5.3.2 PseudoReasoner Training

The objective function of KG-BERT on the labeled dataset is shown in Equation (5.1), and the objective function on pseudo labels can be written as:

$$J_U(\theta) = \mathbb{E}_{x_u \sim P_U(x)} \mathbb{E}_{\hat{y} \sim q(y|x_u)} [-\log P_\theta(\hat{y}|x_u)]. \tag{5.6}$$

Here P_L and P_U are the distribution of the labeled and unlabeled dataset, respectively. $q(y|x)$ is the distribution of pseudo labels, modeled by the teacher model and filters. After finetuning KG-BERT initialized with θ^* on the filtered pseudo labels with Equation (5.6), we acquire $\theta^{*\prime}$.

5.4 Experiments

5.4.1 Baselines

For the supervised learning setting, we use KG-BERT [4] and COMET [56] (GPT2) to perform CSKB Population. For KG-BERT, as it's flexible to be adapted using different pre-trained encoders, we use BERT [35], RoBERTa [175], DeBERTa [176], and BART [177] as the backbone language models. For BERT, RoBERTa, and DeBERTa, we use the embedding of the [CLS] token in KG-BERT as the representation of the whole triple. For BART, we follow the ways of doing sequence classification in the original paper [177] to use the embedding of the end-of-sentence token in the decoder as the representation of the whole triple.

For the semi-supervised learning setting, we use the following baseline models:

Unsupervised Data Augmentation (UDA). UDA [5] uses consistency training to constrain the model to provide invariant predictions with noise added to the input. We adapt

UDA into the framework of CSKB Population and uses TF-IDF word replacement and back-translation to provide noise to the input text to be fed into the consistency loss.

Noisy Student. Noisy student [7] trains a student model with noise added during training iteratively. A teacher model is first trained to provide hard or soft pseudo labels for a student model to finetune together with the labeled dataset. Soft pseudo labels mean using logit scores after softmax as labels. Then the student model is iteratively re-used as the teacher model and a new student model is acquired through each iteration.

Generative Data Augmentation (G-DAUG). G-DAUG [6] leverages text generation language models to automatically generate pseudo training data examples for finetuning. Though it is not designed for semi-supervised learning, we adapt it to our framework of pseudo labeling to serve as a semi-supervised learning baseline. We use COMET (GPT2-XL) finetuned on D_l to generate pseudo examples with heads from D_u. Then, those pseudo labels are filtered with influence function, diversity heuristics, and KG-BERT scores. Then, those pseudo examples are used in the same way as in our PseudoReasoner for further finetuning. We also try to replace COMET with COMET-distill [178], the COMET trained with distilled commonsense knowledge from GPT3, which has a better performance and capacity than the vanilla COMET.

5.4.2 Experimental Settings

The learning rate for all models are set as 1e-5, and the batch size is 64. We use the framework of Huggingface Transformers[1] to form our codebase. Early stopping is used when the best checkpoint is selected and the largest validation AUC is achieved. For all experiments, we report the average scores across three different random seeds.

For thresholding, we set the thresholds $\mathcal{T}^-_{min}=-4.0, \mathcal{T}^-_{max}=-3.7, \mathcal{T}^+_{min}=-2.8, \mathcal{T}^+_{max}=-2.0$, by roughly observing the data distribution and representative knowledge triples in different range of plausibility scores. We then randomly down-sample the pseudo examples so that the number is the same as the original training data. Ablations are provided in section 5.5.1

[1] https://huggingface.co/

Teacher models	all	*Original Test Set*	*CSKB head + ASER tail*	*ASER edges*
N/A (Baseline)	70.9	78.0	63.4	64.6
w/o all filters				
RoBERTa-large	71.8	79.1	65.1	64.6
GPT2-small	72.3	79.5	65.1	65.9
GPT2-medium	72.6	79.5	65.8	68.0
GPT2-XL	72.8	80.1	66.3	66.0
w/ all filters				
RoBERTa-large	72.3	79.2	65.4	66.3
GPT2-small	73.7	79.8	67.5	68.8
GPT2-medium	74.1	**81.3**	68.1	69.0
GPT2-XL	**74.2**	80.1	**69.5**	**69.3**

Table 5.4: Performance of **PseudoReasoner** (RoBERTa-large) using different teacher models w/ or w/o filters.

5.4.3 Results

The main results are shown in Table 5.3. We compare the results of both supervised learning and semi-supervised learning approaches with our proposed model PseudoReasoner. The "all" column presents the overall AUC across all testing examples and is the main metric of CSKB Population. We also separately present the AUC of different test set categories, In-domain (*Original Test Set*) and Out-of-domain (*CSKB head + ASER tail*, and *ASER edges*). In the last two columns, we report the increase by applying different semi-supervised learning approaches under the same backbone model, where Δ_{all} means the increase of "all" (AUC) metric and Δ_{OOD} means the increase of AUC for out-of-domain test sets.

For supervised-learning approaches, KG-BERT-based models mostly perform well on the In-domain test set while have poorer generalization ability to Out-of-domain test sets compared to COMET (GPT2). As GPT2 is only finetuned on the positive part of the dataset, it suffers less from the bias of negative sampling in D_l^- and has a better generalization on new knowledge. However, the drawback in In-domain reasoning hinders the overall performance of GPT2-XL from surpassing KG-BERT (RoBERTa-large), even with 4.5 times more parameters.

Semi-supervised learning baselines can increase the performance of the backbone KG-BERT (RoBERTa-large), especially for the out-of-domain split. However, the improvement

Filter	all	*Original Test Set*	*CSKB head + ASER tail*	*ASER edges*
baseline	70.9	78.0	63.4	64.6
w/o filter	72.8	80.1	66.3	66.0
+ influence	73.2	78.0	68.3	**70.6**
+ KG-BERT	73.7	79.5	68.9	68.6
+ both	**74.2**	**80.1**	**69.5**	69.3

Table 5.5: The effects of different filters on pseudo labels when KG-BERT (RoBERTa-large) is the backbone, and GPT2-XL is the teacher model.

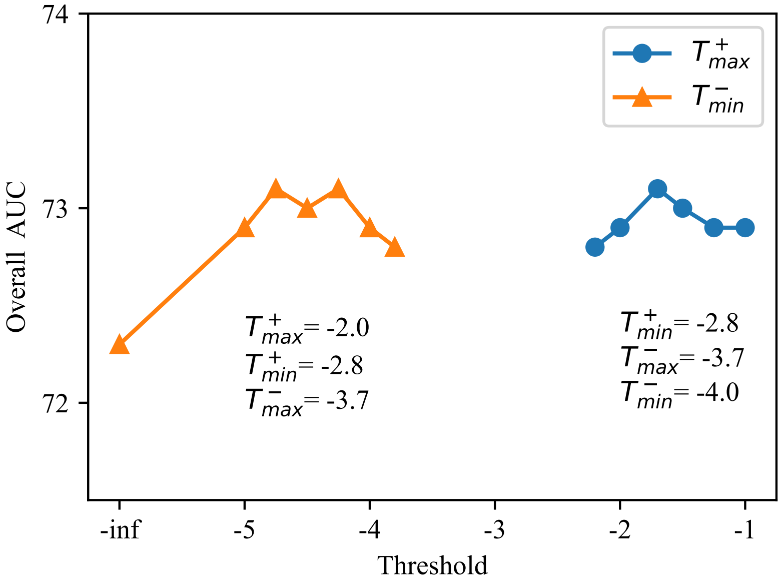


Figure 5.3: Ablation study on different $\mathcal{T}^+_{max}$ and $\mathcal{T}^-_{min}$.

on the In-domain part remains insignificant, and the improvement on the out-of-domain part is not as competitive as our PseudoReasoner. As we use the same code base and training method to train all semi-supervised learning methods, the main differences between PseudoReasoner and other SSL methods lie in the ways of processing the unlabeled dataset. We leave the detailed discussions in the next section (Section 5.5.2).

5.5 Analysis and Discussions

In this section, we discuss the ablation study on model components, the comparisons with semi-supervised learning baselines, diversity analysis, and discussions about why PseudoReasoner works.

5.5.1 Ablation Study

We study the effects of different teacher models and the filters on pseudo labels.

Teacher Models

We compare four representative teacher models, KG-BERT (RoBERTa-large) and GPT2 (small, medium, and XL), on how pseudo labels provided by them can influence the model performance. The ablation results are shown in Table 5.4. From the comparison between KG-BERT (RoBERTa-large, 340M parameters) and GPT2-small (117M) and -medium (345M), which are of the same scale of model size, we find that GPT2 can perform consistently better than KG-BERT as teacher models. This can be validated by the out-of-distribution performance in Table 5.3 for the supervised learning baselines, where KG-BERT performs almost 3 points behind GPT2-medium in terms of OOD AUC. This ablation indicates the importance of powerful generalizable teacher models on pseudo-labeling.

Thresholding

We study the sensitivity of thresholds in Figure 5.3. In this ablation, for simplicity, we set different $\mathcal{T}^{+}_{max}$ for positive pseudo examples, and sample the same amount of triples as the original training set whose $\alpha(x) < \mathcal{T}^{+}_{max}$ in descending order. We do the same ablation study on $\mathcal{T}^{-}_{min}$ for negative pseudo examples. When tuning one threshold, other thresholds are fixed as in section 5.4.2. The pseudo labels under different thresholds are directly used for PseudoReasoner without filtering, and we plot the test set AUC given different thresholds. We see that the resulting AUC is stable within certain ranges of $\mathcal{T}^{+}_{max}$ and $\mathcal{T}^{-}_{min}$. While when we set $\mathcal{T}^{-}_{min}$ to $-\infty$, which indicates no thresholds for negative examples are set, the performance drops drastically.

Pseudo Label Filter

We conduct experiments with different combinations of filtering mechanisms in Table 5.5 for KG-BERT (RoBERTa-large). We can see that both filters (influence function and KG-

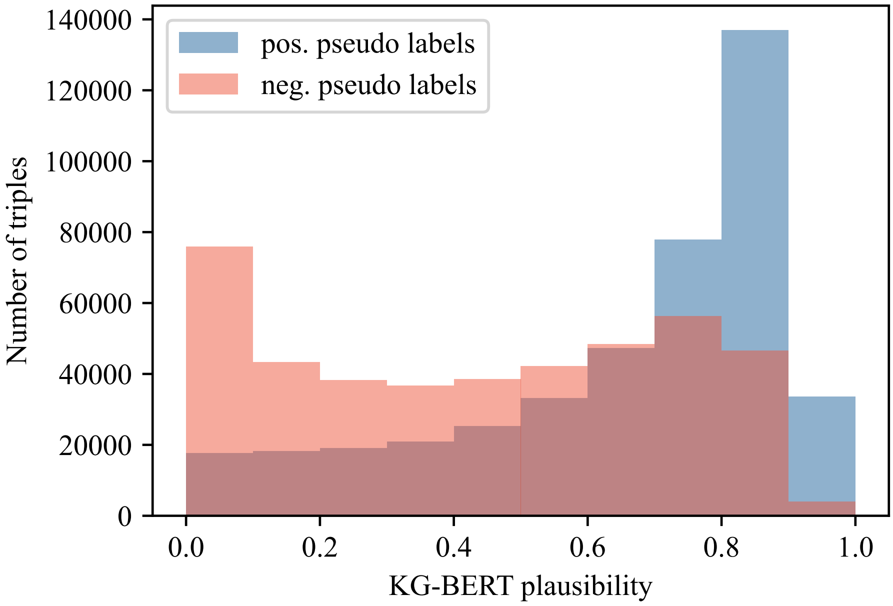


Figure 5.4: KG-BERT plausibility distribution for positive/negative pseudo labels provided by GPT2.

	w/o filter	influence	KG-BERT	both
UDA	3.3 M	-	-	-
G-DAUG	1.1M	399.8K	323.3K	160.1K
PseudoReasoner	932.6K	408.5K	373.0K	170.1K
Original	1.1 M	-	-	-

Table 5.6: Number of pseudo examples used in experiments for semi-supervised methods. The "Original" row indicates the number of training examples in the original training set.

BERT probability) benefit the model performance, while KG-BERT probability contributes to a more substantial improvement. Figure 5.4 shows an illustration of the KG-BERT plausibility $\alpha(x) = P^*_\theta(y{=}1|x)$ of positive/negative pseudo examples provided by GPT2. The positive pseudo examples tend to score higher on KG-BERT than negative pseudo examples. Adding KG-BERT probability as an additional filter with the labels provided by GPT2 is similar to an ensembling procedure.

5.5.2 Comparisons with Other Semi-supervised Learning Methods

Computational Cost

The number of pseudo examples used for semi-supervised-learning baselines are listed in Table 5.6. We basically use the same scale of unfiltered pseudo examples for G-DAUG and

PseudoReasoner while using 3 times more unlabeled examples for UDA as it requires more unlabeled data. Specifically, G-DAUG (COMET-distill) leverages the distilled knowledge from GPT3 [178], which further magnifies the computational cost to more orders of magnitude. In all, under the same scale of pseudo labels, PseudoReasoner can achieve far better results than UDA and G-DAUG.

Analysis

In UDA, though robustness can be improved with consistency loss on noised inputs, there is no *new* commonsense knowledge added to the training procedure, making it hard for the model to be equipped with novel knowledge reasoning ability.

For G-DAUG, the critial part lies in the generation of negative examples. We finetune two separate GPT2 on D_l^+ and D_l^-, and the one finetuned on D_l^- is used to generate negative examples. Compared with the GPT2 finetuned on D_l^+, the GPT2 finetuned on D_l^- is of a relatively lower quality as triples in D_l^- don't follow specific commonsense patterns. We check the text generation quality of GPT2-XL finetuned on D_l^+ and D_l^- and find that the BLEU-2 scores on two corresponding held-out test sets are 0.23 and 0.10, indicating the generation of negative examples are of lower quality.

For Noisy-student, the main differences between our PseudoReasoner is that they use KG-BERT to provide pseudo labels, and they are soft pseudo labels. The main reason behind is that as soft labels are used, the teacher model has to be a discriminative model such as KG-BERT, which has a poor generalization ability than GPT2 used in our PseudoReasoner. Moreover, similar to the case in UDA, noisy is not a dominate factor in CSKB Population, while more high-quality novel commonsense knowledge matters more.

5.5.3 Semantic Diversity Analysis

An important contribution of PseudoReasoner is that we extend the knowledge space for training from limited CSKBs to a more broad, unlabeled resource. We use the proportion of unique uni-grams and bi-grams as an indicator of semantic diversity to measure the scale to which models are exposed to diverse novel knowledge. Figure 5.5 shows that after filtering

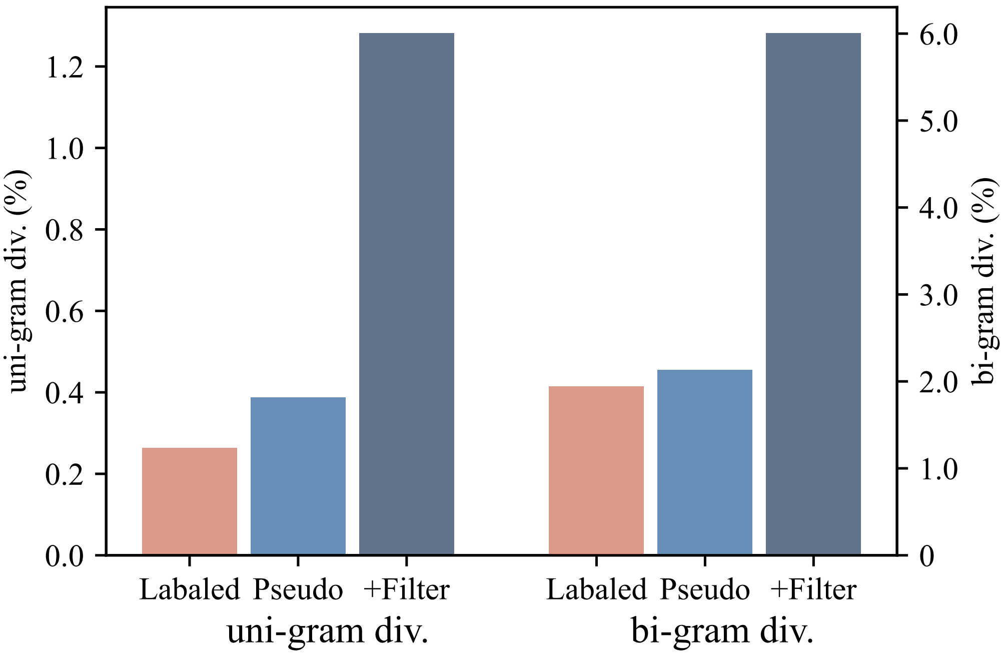


Figure 5.5: Diversity analysis with the proportion of unique uni/bi-grams in the labeled dataset, the pseudo labels, and the filtered pseudo labels. With filtering, the diversity can be significantly improved.

with influence function and KG-BERT probability, the diversity can be improved by around 3 to 4 times than the labeled dataset.

5.5.4 Relationship with Knowledge Distillation

While knowledge distillation focuses on distilling knowledge from larger models to smaller ones, our method does not necessarily need the teacher model to be more significant. The teacher GPT2-medium, which is of the same size as the student KG-BERT, can work pretty well and is comparable to GPT2-XL.

5.6 Conclusion

In this section, we propose a semi-supervised learning framework for CSKB Population based on pseudo labels. Using a teacher model and a special filtering mechanism on pseudo labels, we achieve the state-of-the-art of CSKB Population in terms of both in-domain and out-of-domain performance. Experiments also show that our CSKB Population benefits more from high-quality novel knowledge than other semi-supervised learning techniques such as noise and consistency training. This work brings a new perspective of improving out-of-domain generalizable commonsense reasoning ability on CSKBs.

CHAPTER 6

LEVERAGING COMMONSENSE KNOWLEDGE BASES FOR REASONING

6.1 Extrinsic Evaluation

In this section, we study two downstream applications of CKBP. After acquiring a population model, it acts as a scoring function to determine whether a triple from the candidate knowledge base G is plausible or not, thus serving as a source of commonsense knowledge acquisition [110]. We leverage the populated knowledge as additional training data for both generative commonsense inference (COMET; [56]) and zero-shot commonsense question answering [162].

6.1.1 Generative Commonsense Inference (COMET)

Setup We follow the basic settings as in the original ATOMIC$_{20}^{20}$ paper [10] to generate commonsense tails t given head h and relation r as input. The evaluation dataset is the annotated 5,000 test examples provided by ATOMIC$_{20}^{20}$ [10]. We use BLEU [179], ROUGE-L [180], METEOR [181], and CIDEr [182] as the automatic evaluation metrics.

Specifically, we compare the performance of the following training paradigms: 1) Training the model using the official training set of ATOMIC$_{20}^{20}$. 2) Pre-training the model using a comparable amount of CKBP-acquired data, and subsequently fine-tune on ATOMIC$_{20}^{20}$ training set. 3) Training on a mixture of CKBP-acquired data and ATOMIC$_{20}^{20}$ training data.

We filter the CKBP-acquired data using two filters. First, we employ two typical population models, RoBERTa-L [163] fine-tuned on CKBP training set and Vera [161] to provide a plausibility score for each triple. We set an empirical threshold of 0.8 and selected triples with plausibility scores higher than that of populated commonsense knowledge. Second, we utilize a diversity filter defined in G-DAUG [183], which is a heuristic favoring diverse

Training Data	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
ATOMIC	41.8	26.6	19.2	14.5	50.0	21.2	66.1
ATOMIC + CKBP$_{\text{RoBERTa-L (V1)}}$	41.9	26.6	18.8	13.8	49.7	21.2	66.2
ATOMIC + CKBP$_{\text{RoBERTa-L (V2)}}$	42.5	26.7	18.8	13.8	50.2	21.4	67.1
ATOMIC + CKBP$_{\text{vera}}$	42.9	27.2	19.4	14.4	50.2	21.4	**67.5**
ATOMIC + CKBP$_{\text{vera (mix)}}$	**43.3**	**27.6**	**19.7**	**14.7**	**50.3**	**21.5**	67.4

Table 6.1: Performance (%) of GPT2-Large on generative commonsense inference modeling (COMET). B is for BLEU scores. ATOMIC stands for ATOMIC$^{20}_{20}$ training set, and CKBP stands for our CKBP data. Subscripts under CKBP indicate the population model to select populated commonsense knowledge. The best performances are **bold-faced**.

n-grams. The diversity filter is applied such that we select the same amount of CKBP-acquired data as the training set of ATOMIC$^{20}_{20}$.

We choose GPT2-Large as our backbone language model. We didn't use GPT2-XL as in [10] because the XL version performs relatively poorer than the Large version in terms of most automatic evaluation metrics on the evaluation set of ATOMIC$^{20}_{20}$ despite twice the model size. The learning rate is set as 1e-5, and we train the model for three epochs on both CKBP-acquired data and ATOMIC$^{20}_{20}$ training data.

Results and Analysis The results of generative commonsense inference are presented in Table 7.3. First, adding CKBP-acquired commonsense knowledge for either pre-training or co-training can yield a general performance improvement in generative commonsense inference. Specifically, the model trained on ATOMIC + CKBP Vera achieves the best performance and outperforms that only fine-tuned on ATOMIC$^{20}_{20}$ on all automatic evaluation metrics. This indicates that leveraging the abundant unlabeled discourse knowledge from ASER (G), accompanied by appropriate plausibility filtering through the population model, can effectively serve as valuable augmented data to enhance commonsense reasoning. Among the population models, we observe that a better population model, as evaluated by our CKBP v2 evaluation set, corresponds to a higher performance gain in the generative commonsense inference task. This finding highlights the promising potential of developing improved population models, which subsequently contribute to enhanced downstream applications.

Second, the RoBERTa-L model selected by CKBP v2 demonstrates greater efficacy in enhancing generative commonsense inference compared to the model selected by CKBP

Model	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Zero-shot Baselines							
Random	-	50.0	20.0	50.0	33.3	50.0	40.7
Majority	-	50.8	20.9	50.5	33.6	50.4	41.2
RoBERTa-L [163]	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L [186]	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk [169]	-	-	32.4	70.2	46.2	54.7	-
COMET-DynGen [170]	ATOMIC	-	-	-	50.1	-	-
SMLM [187]	*	65.3	38.8	-	48.5	-	-
MICO [188]	ATOMIC	-	44.2	-	56.0	-	-
STL-Adapter [189]	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
Backbone: DeBERTa-v3-Large *435M*							
DeBERTa-v3-L (MR) [162]	ATM-10X	75.1	71.6	**79.0**	59.7	71.7	71.4
DeBERTa-v3-L (MR) [162]	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8
DeBERTa-v3-L (MR) [162]	CKBP (our)	**79.2**	69.6	77.9	64.3	**77.2**	**73.6**
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	69.3	**74.5**	75.1	**69.5**	62.8	70.2
Supervised Learning & Human Performance							
RoBERTa-L (Supervised)	-	85.6	78.5	79.2	76.6	79.3	79.8
DeBERTa-v3-L (Supervised)	-	89.0	82.1	84.5	80.1	84.1	84.0
Human Performance	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 6.2: Zero-shot evaluation results (%) on five commonsense question answering benchmarks. The best results are **bold-faced**, and the second-best ones are underlined. The performance of supervised learning and human are for reference only.

v1. This finding suggests that CKBP v2 exhibits improved alignment with real-world downstream applications, surpassing its predecessor in terms of practical utility. It's also noteworthy that COMET is an important task that inherently benefits a pile of further downstream tasks that requires commonsense reasoning, including zero-shot commonsense question answering with self-talk [169] and dynamic graph construction [170], narrative reasoning [184], and dialogue generation [185]. In this regard, our work exhibits significant potential for generalization to tasks extending beyond the realm of commonsense reasoning.

6.1.2 Zero-shot Commonsense QA

Setup For the zero-shot commonsense question answering (QA) task, we adopt the task definition and evaluation pipeline proposed by [162] to evaluate the benefit CKBP v2 brings to extrinsic QA. Several methods have been proposed to tackle this task, including those by [169, 170, 189] The most effective pipeline, as proposed by [162], injects commonsense knowledge into pre-trained language models through fine-tuning on QA pairs synthesized from knowledge in CSKBs. To perform this fine-tuning, the head h and relation r of a (h, r, t) triple are transformed into a question using natural language prompts, while the tail t is used as the correct answer option. Distractors or negative examples are created by randomly sampling tails from triples that do not share common keywords with the head. This fine-tuning process enhances the model's knowledge not only for QA benchmarks constructed from CSKBs, such as SocialIQA [21] derived from ATOMIC, but also improves its ability to answer previously unseen commonsense questions in a more generalized manner.

We adopt the original QA synthesis and model training pipeline by [162] on the original ATOMIC and the one augmented with populated knowledge from CKBP v2 to ablatively study the sole benefit that knowledge in CKBP v2 brings. Similar with that in COMET experiments, we use the best-performed CKBP model, Vera, to score the whole population space in ASER and select the populated knowledge with plausibility scores of over 0.8. Then, the same diversity filter as in Section 6.1.1 is used to downsample the number of populated triples to be comparable with the size of the training set in ATOMIC$^{20}_{20}$. For the QA model, DeBERTa-v3-Large [186] is used as the backbone, and we train the model using a learning rate of 7e-6 for one epoch on both the CKBP-acquired data and ATOMIC-synthesized data as provided by [162].

Once trained, we evaluate the model on the validation splits of five commonsense QA benchmarks: Abductive NLI (aNLI; [190]), CommonsenseQA (CSQA; [17]), PhysicalIQA (PIQA; [125]), SocialIQA (SIQA; [21]), and WinoGrande (WG; [19]). Accuracy is used as the evaluation metric. Furthermore, we compare our model not only against existing zero-shot knowledge injection methods [169, 170, 187, 188, 189, 162] but also against large language models such as ChatGPT [41] and GPT-3.5 [100].

Results and Analysis The zero-shot commonsense QA results are shown in Table 6.2. Among all the zero-shot methods, the model trained on CKBP v2 demonstrates the highest performance. It outperforms models trained solely on ATOMIC (with an increase of 2.2%) and ATOMIC10X [178] (with an increase of 1.8%). Importantly, our method surpasses large language models by an average of 3.4%. This performance gain highlights the significant advantage of our populated commonsense knowledge over both human annotations and distilled knowledge from large language models. Furthermore, we observe that the model trained on CKBP-acquired data shows the most improvement on the aNLI and WinoGrande benchmarks. One potential reason for this is that the populated knowledge in CKBP v1 encompasses a wider range of commonsense knowledge beyond only social commonsense, which benefits tasks involving abductive reasoning (based on narrative) and pronoun coreference resolution.

6.2 Conclusion

We propose the pipeline of using populated commonsense knowledge as training data for 1) question answering models and 2) generative commonsense inference models. Experiments show that including the populated knowledge as additional training data can help improve downstream commonsense reasoning on six datasets.

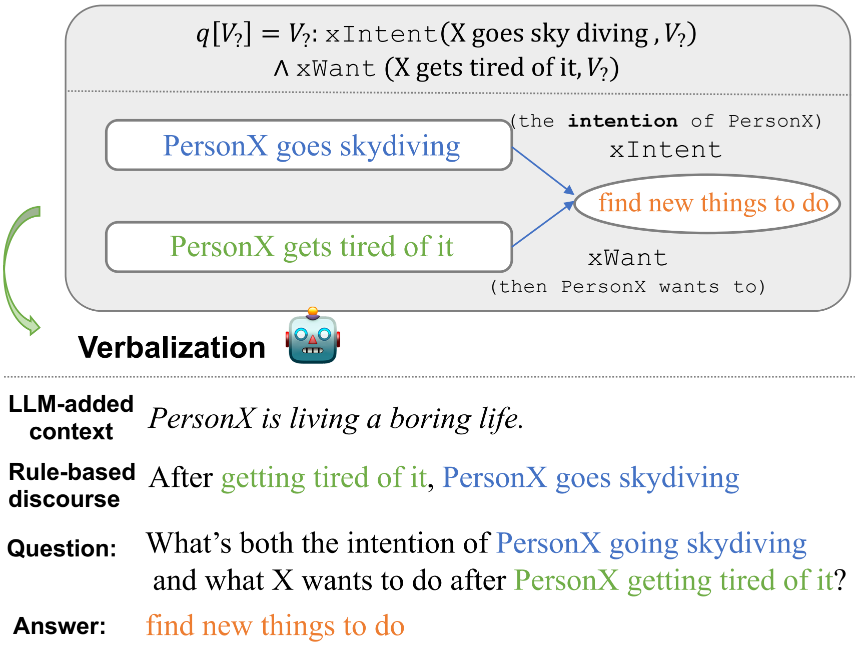
CHAPTER 7

COMPLEX REASONING OVER COMMONSENSE KNOWLEDGE BASES

7.1 Introduction

Large language models struggle to effectively perform reasoning when presented with complex tasks, such as reasoning about multiple events and their relationships. This shortcoming is due to both the inherent difficulty of reasoning over multiple pieces of information, as well as a lack of adequate-scale, supervised training datasets for learning [61]. Unfortunately, complex and multi-hop commonsense reasoning benchmarks [9] are both technically challenging and financially expensive to curate. Consequently, previous efforts either constructed datasets (a) with simpler reasoning structures, such as single-hop chains [62], (b) using distant supervision based on one-hop inference [9], or (c) with human-annotations, but at a relatively small scale [8].

To alleviate this training data bottleneck, recent works have explored extracting and formulating questions from existing CommonSense Knowledge Graphs (CSKGs [10]), which store commonsense triples. However, using CSKGs to produce high-quality reasoning datasets poses several challenges. First, while the shared entities in commonsense triples encode a complex, interconnected graph structure, the sparsity of this structure limits the number of potential questions that encode more than one reasoning hop [21, 49]. Second, triples in CSKGs are represented in a context-free manner, such as the event “PersonX gets tired of it” in Figure 7.1, yielding ambiguous (and sometimes incorrect) human annotations in the CSKG, e.g., ATOMIC [191] has an error rate of over 10%. These errors propagate when triples are naively combined to construct reasoning questions. Finally, also because triples in CSKGs are represented in a context-free manner, additional context must be added to make questions fluent, a problem exacerbated in multi-hop settings where the entities of multiple reasoning hops must be coherently verbalized together.


$q[V_?] = V_?$: xIntent(X goes sky diving , $V_?$)
∧ xWant (X gets tired of it, $V_?$)
PersonX goes skydiving
PersonX gets tired of it
(the **intention** of PersonX)
xIntent
find new things to do
xWant
(then PersonX wants to)
Verbalization

LLM-added context *PersonX is living a boring life.*

Rule-based discourse After getting tired of it, PersonX goes skydiving

Question: What's both the intention of PersonX going skydiving and what X wants to do after PersonX getting tired of it?

Answer: find new things to do

Figure 7.1: An example of conjunctive logical queries and their verbalization as complex commonsense inferences.

In this section, we construct COM2 (**COM**plex **COM**monsense), a novel commonsense reasoning dataset using multi-hop queries in commonsense knowledge graphs to construct question-answer pairs requiring complex narrative reasoning. To build this dataset, we use *conjunctive logical queries* [63], a subset of First-Order Logical queries that use existential quantifiers and conjunction. The multi-hop projection operation involves inferring hidden contexts, while the intersection operation enables reasoning among multiple events, encompassing common cause or effect, and abduction. For example, in Figure 7.1, an intersection of two triples can be verbalized to a short narrative, and the process of inferring the common tail can be seen as an *abduction* of the hidden cause between the two heads.

To address the challenges above, we propose to first *densify* the CSKG to merge nodes with high semantic similarity, increasing the connectivity of the graph. Then, we use an off-the-shelf plausibility scorer to filter out low-quality triples, avoiding error propagation as we construct more complicated queries. Finally, we verbalize the queries in a natural language context with handcrafted rules and Large Language Models to derive coherent and informative narrative contexts for our questions. Our final COM2 dataset comprises 790K question-answer pairs (both with multiple-choice and generative answer settings), including 1.3K examples that we manually verify for evaluation.

Our results demonstrate the challenges faced by even powerful LLMs and supervised question-answering models on the COM2 dataset, underscoring the difficulty of performing complex multi-hop reasoning. Moreover, fine-tuning question-answering models and generative commonsense inference models on COM2 leads to substantial improvements across eight commonsense reasoning datasets, showing the efficacy of our framework for boosting commonsense reasoning ability.

To conclude, our contributions are three-fold. First, we present a pipeline for sampling and verbalizing complex logical queries from CSKGs, to form a complex commonsense reasoning benchmark, COM2, with minimal human effort. Second, we benchmark the complex reasoning ability of various state-of-the-art language models and question-answering models on COM2. Finally, we validate the benefit of fine-tuning COM2 on eight zero-shot commonsense reasoning datasets.

7.2 Methodology

In this section, we introduce the construction details of COM2, including the pre-processing, sampling, and verbalization of complex queries, as well as the details of human annotations.

7.2.1 Pre-processing

We use ATOMIC$^{20}_{20}$ [10], a comprehensive Commonsense Knowledge Graph covering everyday social, physical, and event-level knowledge, as the base CSKG. Before sampling queries, we address the sparsity and quality issues first.

Sparsity CSKGs are usually highly sparse compared to factual KGs due to the diversity and scale of commonsense [192], resulting in many isolated nodes that can hardly be sampled as part of a complex query. To alleviate this issue, we develop a set of rules and use sentence embedding similarity to merge nodes in the CSKG, leading to 22.4% of nodes being merged and an average degree increase of 25.3%.

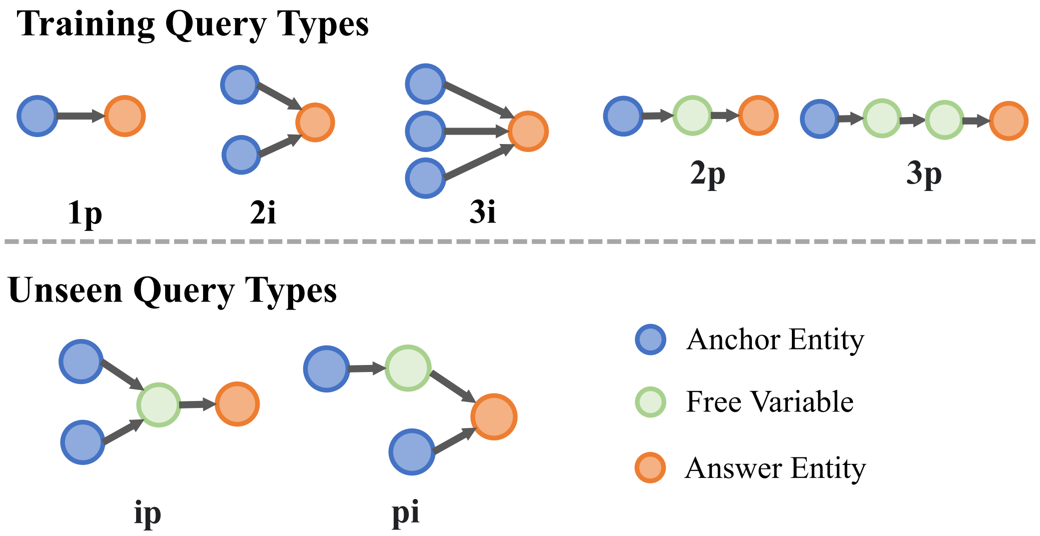


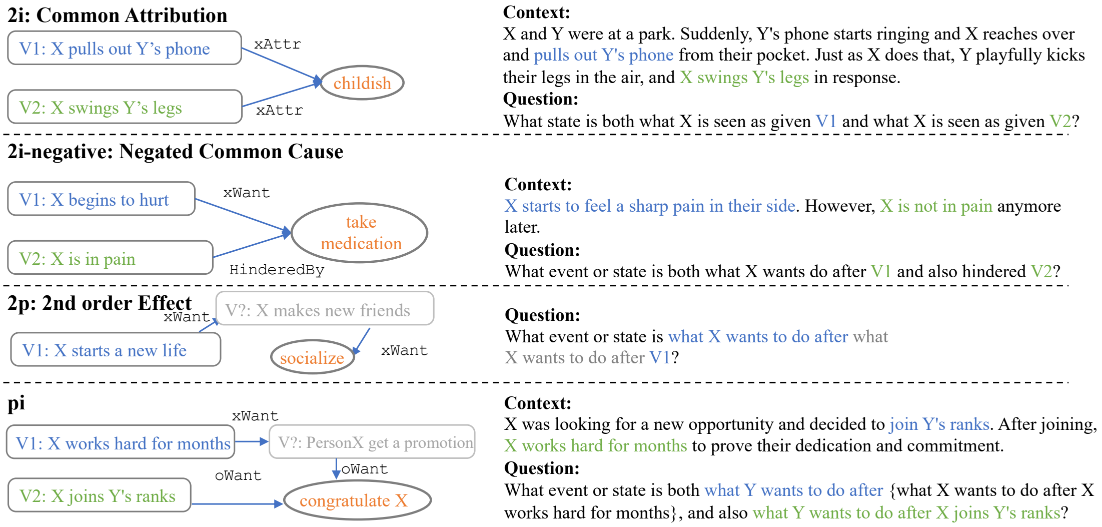
Figure 7.2: Visualization of query structures. The anchor entities and relations are specified to instantiate the query. ‘p’ and ‘i’ represent *projection* and *intersection*, and the number ahead of p and i indicates the number of anchor entities and free variables.

Quality The error rate of CSKGs (e.g., ATOMIC has an error rate of ∼10%) can be problematic when we consider the intersection and projection of more than two triples (errors in a single triple could propagate to many multi-hop queries). We use an off-the-shelf plausibility scorer Vera [193], a T5-based scorer fine-tuned on 2 CSKGs and 19 QA datasets, to score every triple in terms of commonsense plausibility (between 0 to 1). We filter out triples (∼10%) with a plausibility score lower than 0.5, the threshold provided in Vera [193] for plausible statements.

7.2.2 Query Sampling

The query structures that we study are visualized in Figure 7.2. Following Query2box [128], we use projections (1p, 2p) and intersections (2i, 3i) as training queries, and leave complex queries ip and pi as zero-shot evaluation queries. To examine scenarios involving negation and differentiate them from regular 2i queries, we use the term “2i-neg” to represent 2i queries where one of the relations is “HinderedBy”. In this formulation, multi-hop projection involves inferring hidden reasoning contexts, while intersection operations require reasoning about complex interactions between events.

Given a query structure, we use pre-order traversal to sample free variables and anchor entities starting from an answer entity. We sample predecessors uniformly based on (rela-



Context:
X and Y were at a park. Suddenly, Y's phone starts ringing and X reaches over and pulls out Y's phone from their pocket. Just as X does that, Y playfully kicks their legs in the air, and X swings Y's legs in response.
Question:
What state is both what X is seen as given V1 and what X is seen as given V2?

Context:
X starts to feel a sharp pain in their side. However, X is not in pain anymore later.
Question:
What event or state is both what X wants do after V1 and also hindered V2?

Question:
What event or state is what X wants to do after what X wants to do after V1?

Context:
X was looking for a new opportunity and decided to join Y's ranks. After joining, X works hard for months to prove their dedication and commitment.
Question:
What event or state is both what Y wants to do after {what X wants to do after X works hard for months}, and also what Y wants to do after X joins Y's ranks?

Figure 7.3: Examples of different query types, their verbalization, and corresponding questions.

tion, entity) pairs. During sampling, to avoid over-sampling on nodes with extremely high degrees, we empirically set a cut-off degree $\mathcal{T} = 10$ to only sample from the top $\mathcal{T}$ neighbors of a node scored by Vera. In the end, we conduct a post-order traversal starting from the anchor entities to find all the answers to the query, in addition to the starting answer entity.

Distractor Sampling We sample 4 additional candidate distractors for each query, where 2 of them are randomly sampled across the whole CSKG, and 2 of them are sampled from the neighbors of the anchor entities that are not the answers to the whole query, represented as adversarial negative examples. When fine-tuning a question-answering model, the negative examples are used as synthetic question-answering pairs for training. In the evaluation set, these candidate negative examples, together with the sampled answer, are manually annotated to form a gold evaluation set.

7.2.3 Verbalization

CSKGs are constructed in a context-free manner. To make the logical queries on such context-free triples more human-interpretable, we introduce an additional step of verbalizing the anchor entities to a narrative, to effectively acquire fluent and plausible narrative-inference pairs.

Method	2i	2i-neg	3i	2p	ip	pi	All
API-based LLMs							
`gpt-3.5-turbo-0613`	33.56	43.12	42.01	38.66	38.05	28.40	37.74
- 1-shot	43.31	35.31	58.45	57.73	51.33	62.96	48.22
- 1-shot w/ CoT	45.80	36.43	54.34	57.73	50.44	66.67	48.75
- 8-shot (2i, 2p)	48.52	41.26	57.08	67.53	53.10	74.07	53.22
- 8-shot (2i, 2p) w/ CoT	52.61	46.10	60.27	59.79	52.21	65.43	54.37
`gpt-4-1106-preview`	44.67	46.47	52.05	32.47	40.71	53.08	44.64
- 1-shot	47.85	42.01	50.68	38.66	44.25	50.62	45.63
- 1-shot w/ CoT	48.97	46.46	52.96	49.48	52.21	58.02	50.04
- 8-shot (2i, 2p)	54.87	46.47	58.90	45.88	52.21	66.67	53.00
- 8-shot (2i, 2p) w/ CoT	57.82	49.07	62.56	61.34	52.21	66.67	57.40
Open-source (QA) Language Models							
HyKAS (162, zero-shot)	34.92	39.41	27.85	41.75	37.17	33.33	35.76
CAR (194, zero-shot)	37.41	30.48	37.44	57.73	32.74	53.09	39.56
Llama2 (7B) [195]	35.15	21.93	39.27	35.57	28.32	51.85	33.64
Vera (5B) [193]	47.62	27.51	40.18	66.49	52.21	58.02	46.09
UnifiedQA-v2 [196]	56.23	39.41	62.56	58.76	51.33	62.96	54.21
Flan-T5 (11B) [197]	58.28	47.21	65.30	**76.29**	56.64	79.01	60.97
Fine-tuned on COM2							
DeBERTa-v3-Large (+COM2)	60.09	**58.36**	69.41	61.86	**59.29**	81.48	62.79
CAR-DeBERTa-v3-Large (+COM2)	**61.22**	56.13	**69.86**	68.56	56.64	**85.19**	**63.78**

Table 7.1: Model performance (%) on the multiple-choice question answering evaluation set of COM2.

Anchor Entity Verbalization We consider a rule-based verbalizer and a ChatGPT-driven verbalizer. In the rule-based verbalizer, we add a discourse marker between the two or three anchor entities depending on the semantics of the query relations. For example, a simple situation would be adding an “and” or “then” between two anchor entities in a 2i query. To make the query more human-understandable, we consider using ChatGPT to synthesize the necessary contexts to make the query an actual narrative.

Relation Verbalization The multiple relations in complex queries can be deterministically converted to a question using the natural language descriptions of the relations. Examples can be found in Figure 7.3.

Model	CSKG	Out-of-domain						In-dom.
		a-NLI	CSQA	PIQA	SIQA	WG	Avg.	COM2
Random	-	50.0	20.0	50.0	33.3	50.0	40.7	20.0
DeBERTa-v3-L [186]	-	59.9	25.4	44.8	47.8	50.3	45.6	14.7
Self-talk [169]	-	-	32.4	70.2	46.2	54.7	-	-
COMET-DynaGen [170]	ATOMIC	-	-	-	50.1	-	-	-
SMLM [187]	*	65.3	38.8	-	48.5	-	-	-
MICO [188]	ATOMIC	-	44.2	-	56.0	-	-	-
STL-Adapter [189]	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7	-
Large Language Models								
GPT-3.5	-	61.8	68.9	67.8	68.0	60.7	65.4	-
GPT4	-	75.0	43.0	73.0	57.0	77.0	65.0	44.6
ChatGPT	-	69.3	74.5	75.1	69.5	62.8	70.2	37.7
+ zero-shot CoT	-	70.5	**75.5**	79.2	**70.7**	63.6	71.9	28.9
Backbone: DeBERTa-v3-Large *435M*								
HyKAS [162]	ATM-10X	75.1	71.6	79.0	59.7	71.7	71.4	27.7
HyKAS [162]	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8	35.8
CAR [194]	ATOMIC	78.9	67.2	78.6	63.8	78.1	73.3	36.8
CAR [194]	ATM^C	79.6	69.3	78.6	64.0	78.2	73.9	39.8
HyKAS + COM2 (Ours)	ATM, COM2	78.4	69.9	78.7	64.1	**78.3**	73.9	62.8
CAR + COM2 (Ours)	ATM^C, COM2	**81.2**	70.9	**80.3**	65.6	77.4	**75.1**	**63.8**
Human Performance	-	91.4	88.9	94.9	86.9	94.1	91.2	-

Table 7.2: Zero-shot evaluation results (%) on five out-of-domain commonsense question answering benchmarks, and the in-domain evaluation set of COM2. The best results are **bold-faced**, and the second-best ones are underlined.

7.2.4 Human Annotation

To support reliable automatic evaluation, we formalize the problem of complex commonsense reasoning as a multi-choice question answering task, with one true answer, three distractors, and a fifth option indicating "None of the answers are correct". We crowdsourced the answers using Amazon Mechanical Turk (AMT). The workers are given the verbalized query as the context, the verbalized relations as the question, and the sampled (negative) answers. If no sampled answers are correct, then the worker is asked to select an additional "None of the answers are correct" option. If the verbalization itself does not make sense, the worker can also select another option "The context doesn't make sense or is meaningless" and we discard the example. Each question is annotated by three workers. The overall per-option inter-annotator agreement is 78%, and the Fleiss kappa is 0.445, indicating moderate agreement. The workers are paid, on average, 16 USD per hour. Our final dataset consists of ∼782k training examples and 1317 manually validated evaluation examples.

7.3 Experiments

We conducted experiments on the evaluation set of COM2, which was formulated as a multi-choice question answering (MCQA) task. Specifically, we examine the performance of state-of-the-art off-the-shelf language models on COM2 and also study the effect of training a question-answering model on the distantly supervised training set of COM2.

7.3.1 Setup

We use popular API-based and open-source LLMs as baselines. Following the standard practice of prompting LLMs for QA [198], we initialize a prompt that takes "[Context] [Question] [Options]" as the input and ask the model to only output the associated symbol (e.g., 'A') in the QA pair as the prediction. For open-source language models like Flan-T5 and Llama2, we use the same prompt and compute the logits received by each of the options in the first prediction token.

We also study the effect of fine-tuning a question-answering model on the synthetic training queries discussed in Section 7.2.2. We follow the pipeline by HyKAS [162], which fine-tunes language models on QA pairs synthesized from one-hop knowledge in CSKGs and extends them to complex queries. For one-hop (1p) triples, the head and relation are transformed into a question with pre-defined prompts. For complex queries, the verbalized queries (as illustrated in Section 7.2.3) are regarded as the context, and questions are also transformed with a different prompt template depending on the relations. The tails to the one-hop triple or the sampled answer to the query are regarded as the correct answer, and the negative examples are randomly sampled across the whole CSKG following a keyword overlapping filtering [162, 194]. We use DeBERTa-v3-large as the backbone encoder.

7.3.2 Results and Analysis

Our results are presented in Table 7.1. We observe that Chain-of-Thought (CoT) improves reasoning performance, as it allows the model to first induce the causes or effects of individual events in intersection-based queries (2i and 3i), or induce hidden variables in projection-

Model	Training Data	Multi-Event			Paragraph-Level			Single-Event			COM2		
		B-2	R-L	BERT	R-L	CIDE	BERT	R-L	CIDE	BERT	R-L	CIDE	BERT
(Distantly) Supervised Learning													
COMET-M (BART-L)	MEI	25.1	33.6	64.9	-	-	-	-	-	-	-	-	-
COMET-M (GPT-2-L)	MEI	16.2	25.7	55.1	-	-	-	-	-	-	-	-	-
ParaCOMET (GPT-2-L)	PCD	-	-	-	18.8	27.8	60.2	-	-	-	-	-	-
Zero-shot Learning								**Supervised**					
COMET	1p	1.20	2.73	38.9	3.5	6.4	25.7	50.0	66.1	75.1	10.0	20.7	44.3
COMET-distill	ATM10x	1.20	3.55	12.7	11.8	16.8	29.5	1.6	4.8	24.3	8.3	11.9	36.1
COM2-COMET	1p, 2i	**8.87**	**15.2**	**46.4**	**13.8**	**22.1**	**53.7**	**50.7**	**68.0**	**77.1**	13.6	26.1	39.8
COM2-COMET	1p, 2p, 2i, 3i	5.41	10.4	44.8	9.2	16.6	44.1	50.4	66.9	77.1	**14.7**	**33.0**	**46.3**
LLama2-7b	-	1.81	4.14	45.7	2.2	2.2	48.6	5.4	2.9	51.5	3.9	6.7	44.9
COMET-LLama2-7b	1p	7.62	14.4	44.2	9.1	12.3	51.0	27.5	26.4	64.2	10.9	22.3	44.9
COM2-LLama2-7b	1p, 2i	**8.82**	**16.4**	**47.5**	14.6	**22.1**	55.3	**31.6**	**31.1**	**66.0**	**35.7**	**107.2**	**61.3**
COM2-LLama2-7b	1p, 2p, 2i, 3i	8.22	15.4	47.0	**15.9**	21.3	**55.3**	31.3	29.8	65.5	35.6	105.0	60.1

Table 7.3: Experimental results on downstream narrative commonsense reasoning, including in a multi-event [8] setting, and a paragraph-level setting [9]. In-domain settings include single-event generation and complex inference in COM2. We use BLEU-2 (B-2), ROUGE-L (R-L), CIDEr (CIDE), and BERTScore (BERT) as the evaluation metrics.

based queries (2p as in Figure 7.2). Adding eight-shot exemplars (consisting of 2i, 2i-neg, and 2p queries) further improves performance among prompting baselines.

For models fine-tuned on complex queries using HyKAS and CAR, we observe that the synthetic training pairs, despite lacking manual annotation, serve as valuable distant supervision signals. They enhance the complex reasoning capability of HyKAS and CAR, surpassing the performance of the 8-shot GPT-4 model with CoT by 6%. CAR + COM2 also outperforms the 11B version of UnifiedQA-v2 and Flan-T5, which are both fine-tuned on numerous (commonsense) question answering datasets, by 9% and 3%, respectively.

7.4 Downstream Evaluation

In addition to benchmarking Complex Commonsense Reasoning, we also study the effect of leveraging COM2 as training data to generalize to other downstream commonsense reasoning tasks. As tasks, we use zero-shot CommonSense Question Answering (CSQA), and Generative Commonsense Inference, including one-hop, multi-event, and paragraph-level settings.

7.4.1 Commonsense Question Answering

Setup The task of zero-shot commonsense QA involves selecting the most plausible option for commonsense questions without training on examples from the benchmark dataset. We directly leverage the model we trained in Section 7.3, the DeBERTa-v3-large-based model fine-tuned on synthetic question pairs from both ATOMIC and COM2, and check the performance on five popular commonsense question answering datasets: Abductive NLI (aNLI; [190]), CommonsenseQA (CSQA; [17]), PhysicalIQA (PIQA; [125]), SocialIQA (SIQA; [21]), and WinoGrande (WG; [19]). As baselines, we consider the same methods, HyKAS [162] and CAR [194], but use other CSKGs as training sets. In Table 7.2, ATM-10X refers to ATOMIC-10x from [178], and ATM^C refers to the training data from CAR [194] which is augmented from ATOMIC with conceptualization.

Results and Analysis We report model performance in Table 7.2. We observe the inclusion of COM2 and one-hop triples from ATOMIC as training data for CAR and HyKAS yields significant improvements in question answering ability. Notably, the combination of CAR and COM2 achieves the highest performance among all models, surpassing even ChatGPT and GPT-4, despite having a parameter size of at least two orders of magnitude smaller.

Notably, when using CAR as the base model, training on COM2 leads to the highest performance gain of around 1.8% for a-NLI. When evaluating on a-NLI, which includes instances of abductive reasoning, the model may be helped by learning from 2i queries where one relation represents *cause* and the other represents *effect* (abduction examples in Figure 7.1 and Figure 7.3). Meanwhile, the performance on WinoGrande was adversely affected, likely because Winogrande primarily focuses on identifying distinguishing features of entity pairs. The benefits from learning event-event interactions from COM2 may not transfer well to this setting.

7.4.2 Generative Commonsense Inference

Setup We study generative commonsense inference as an additional evaluation task. We include multi-event commonsense generation (COMET-M; [8]) and paragraph-level commonsense generation (ParaCOMET; [9]) as two out-of-domain evaluation tasks. We also include the vanilla COMET [56] as an additional in-domain evaluation, which focuses on 1p queries that require generating the tail given head and relation as the input. We also conduct experiments on the generative sub-task of COM2, where verbalized context and question inputs are used to inferences. The annotated ground answer options are used as references.

For the (distantly) supervised learning baselines, we fine-tune GPT-2-large on the annotated multi-event inference dataset (MEI) from COMET-M [8] and distantly labeled PCD dataset from ParaCOMET [9] as a reference. In our zero-shot learning setting, we study the effect of fine-tuning COMET (GPT-2-large) on ATOMIC and different query types of COM2. We also study fine-tuning an LLM, Llama2-7b, by converting triples and queries to an instruction-tuning format, following the prompt template in Section 7.2.3. We leverage the framework of Meditron [199][1] to fine-tune Llama2-7b. We fine-tune on a mixture of different query types as detailed in the **Training Data** column. We present the performance results of models fine-tuned on either the annotated or distantly supervised training set for both tasks as reference benchmarks. Specifically, we use MEI for COMET-M and PCD for ParaCOMET. To ensure diversity and prevent overfitting to common tails, complex queries are selected using an n-gram-based diversity filter [183].

Results and Analysis We present the results in Table 7.3. Compared to models fine-tuned solely on one-hop triples, COMET models fine-tuned on additional complex queries demonstrate enhanced generative commonsense inference capabilities for multi-event and paragraph-level scenarios. When comparing different query types, fine-tuning solely on 2i queries yields the most significant improvement in reasoning capability, likely because 2i queries provide more explicit reasoning signals compared to 2p queries, which can be

[1] https://github.com/epfLLM

ambiguous due to the large candidate space of the hidden event. For example, the average number of answers for 2p queries is 7.93, compared with 1.09 for 2i queries. In addition, the answers to 2i queries exhibit greater diversity than 3i queries, as the CSKG is sparse and provides a limited number of distinct tails for sampling 3i queries compared to 2i queries.

7.5 Analysis & Discussion

7.5.1 Ablation Study

We analyze the impact of various data filters, query types, and verbalization methods on generative inference within COM2. Detailed results can be found in Table 7.4.

Filtering We include two types of filters, a Vera-based plausibility filter and a diversity filter. Evaluating the performance of generative commonsense inferences on COM2, we examine the impact of removing both filters while employing GPT2-Large as the backbone model. Removing the plausibility filter results in a significant performance decline, highlighting its critical role. On the other hand, the diversity filter exhibits a minor positive influence on enhancing performance.

Type of Queries We investigate the impact of training our models on different types of logical queries. The model trained only on 1p and 2p queries does not generalize well to other query types, such as pi and ip, leading to a worse performance than the model trained on all query types. However, according to Table 7.1 and Table 7.3, models trained on only 2i queries generalize better to downstream commonsense reasoning tasks, potentially indicating that multi-event reasoning in most existing commonsense benchmarks focuses on intersection more than projection.

Verbalization We investigate the effect of using a rule-based verbalizer or ChatGPT-enabled verbalizer to generate COM2 contexts. Using ChatGPT-verbalized queries leads to better downstream performance on both PCD and COM2. In COM2, the presence of ChatGPT-verbalization intuitively improves performance since the training context aligns

Model	COM2		
	R-L	**CIDEr**	**BERT**
Filter			
COM2-COMET	14.7	33.0	46.3
- w/o plau. filter	13.0	31.2	42.3
- w/o div. filter	14.4	32.5	45.8
- w/o both filter	12.5	30.3	40.1
Query Types			
COMET (1p)	10.0	20.7	44.3
+ 2i	13.6	26.1	39.8
+ 2p	9.8	19.9	43.4
+ 2i, 3i, 2p	14.7	33.0	46.3
Verbalization			
COM2-COMET	13.6	26.1	39.8
COM2-COMET (V)	14.3	27.1	43.4
COM2-Llama	35.7	107.2	61.3
COM2-Llama (V)	36.2	105.4	61.4

Model	PCD		
	R-L	**CIDEr**	**BERT**
Verbalization			
COM2-COMET	13.8	22.1	53.7
COM2-COMET (V)	14.0	23.2	54.0
COM2-Llama	14.6	22.1	55.3
COM2-Llama (V)	14.8	23.6	55.5

Table 7.4: Ablation studies on filters, type of queries, and using ChatGPT for verbalizing queries (denoted as V).

with the evaluation set's format. On the other hand, the context in the PCD dataset is long and comprised of five sentences. Verbalization not only adds more context to the training but also aligns better with the PCD format.

7.5.2 Error Analysis

We present a human-annotated quality evaluation of the Llama-7b-based model on the generation sub-task of COM2. To ensure diverse coverage of query types, we randomly sampled 60 queries, with ten from each of the six types. Manual inspection revealed a common error where the generated output was partially correct, either providing the answer to one of the triples in an intersection query or only the one-hop answer instead of the two-hop answer in 2-projection (2p) queries. Table 7.5 includes the number of such '1-hop' partially

Model	#Plau.	#1-hop	#False
LLama2-7b	26	2	28
COMET-LLama2-7b	29	8	23
COM2-LLama2-7b (2i)	47	2	11
COM2-LLama2-7b (all)	45	3	12

Table 7.5: Human evaluation results on the generative sub-task in COM2 using Llama2-7b as the backbone. '1-hop' indicates the answer is plausible in terms of only one-hop relations.

correct answers. Our results demonstrate that the zero-shot Llama model already produces 26 out of 60 plausible inferences. Fine-tuning the model on one-hop ATOMIC further increases the number of plausible generations while more frequently generating inferences that are one-hop correct. Moreover, fine-tuning on the synthetic training set of COM2 significantly improves the model's ability to generate complex commonsense inferences and reduces the occurrence of partially correct answers.

We present some error cases in Table 7.6. In general, a common error in both projection and intersection queries is that the generated answer can be only the one-hop answer instead of the correct answer that is multi-hop. For example, in the 2p case, "get a new job" is a direct intention of someone who updates his or her resume. However, the 2p query asks about the intention of the intention, which requires inducing the intention behind "get a new job". In this sense, "to be financially independent" is a more plausible inference. In the case of 2i queries, the error lies in the absence of inferential gaps between the context, where the generated answers become paraphrases of the events rather than being the result by any anchor entity. In the case of ip, a common error for one-hop COMET is the generation of "None" for complex cases, indicating a deficiency in multi-hop reasoning capabilities.

7.6 Conclusion

In this section, we leverage the concept of conjunctive logical queries to create a complex commonsense reasoning dataset derived from CSKGs. The dataset, COM2, comprises a human-annotated evaluation set and a distantly supervised training set without further annotations. Our experiments highlight the challenging nature of complex commonsense reasoning that involves multiple events or multi-hop scenarios, even for advanced language models such as GPT-4. Additionally, we train question-answering models and generative commonsense reasoning models using COM2. The results show significant improvements across eight diverse downstream commonsense reasoning tasks, highlighting the potential of leveraging CSKGs to acquire complex reasoning signals inexpensively without relying on extra human effort.

Type	Context	Question	COMET	COM2-COMET
2p	Ezra updates Ezra's resume (V1)	What event or state is the intention of Ezra before the intention of Ezra before V1?	get a new job ✗ (one-hop correct)	be financially independent ✓
2i-neg	Every day, Benjamin goes to work diligently (V1), never missing a day. They are dedicated and committed to their job. In particular, Benjamin doesn't work hard on it (V2) and instead takes a more relaxed approach, focusing on maintaining a healthy work-life balance.	What event or state is both the effect on Benjamin after Benjamin go to work every day (V1) and also what hindered Benjamin work hard on it (V2)?	Benjamin is sick ? (Not perfect as Benjamin is trying to keep a work-life balance instead of having a sick leave)	Benjamin gets tired from working hard ✓
2i	Chloe is known for being hardworking (V1) and dedicated. As a result, Chloe leads a good life (V2).	What event or state is both the effect on Chloe after Chloe is hardworking (V1) and also what Chloe wants to do after Chloe leads a good life (V2)?	to have a good life ? (No inferential gap)	to have success in life ? (No inferential gap)
ip	After looking for a new car (V1), Lydia is driving to school (V2).	What event or state is what Lydia needed to do before the event that is both what Lydia wants to do after Lydia is looking for a new car (V1), and also what Lydia needed to do before Lydia is driving to school (V2)?	None ✗	take a car for test drive ✓

Table 7.6: Error analysis of generated inferences on the evaluation set of COM2. We present the generations of COMET-Llama-7b and COM2-Llama-7b fine-tuned on all queries.

CHAPTER 8

CONCLUSION AND FUTURE WORKS

8.1 Conclusion

First, we study the feasibility of transferring “cheap” and large-scale discourse knowledge to “expensive” inferential *if-then* commonsense knowledge. Experimental results have shown that the proposed DISCOS framework can retrieve much more novel and diverse *if-then* commonsense knowledge from ASER with high quality comparable with neural text generation models.

Second, we formally benchmark CKBP and provide a human-annotated evaluation set containing 37K examples unifying four popular commonsense knowledge bases. We also propose KG-BERTSAGE to both incorporate the semantic of knowledge triples and the subgraph structure to conduct reasoning, which achieves the best performance among other counterparts. Experimental results also show that the task of reasoning unseen triples outside of the domain of CSKB is a hard task where current models are far away from human performance, which brings challenges to the community for future research.

Third, based on the inherent property of the task CKBP, we propose a pseudo-label based semi-supervised learning method to perform population. Using a teacher model and a special filtering mechanism on pseudo labels, we achieve the state-of-the-art of CSKB Population in terms of both in-domain and out-of-domain performance.

Finally, in terms of complex commonsense reasoning, we leverage the concept of conjunctive logical queries to create a complex commonsense reasoning dataset derived from CSKGs. Our experiments demonstrate the difficulty of complex commonsense reasoning, even for advanced language models like GPT-4, when multiple events or multi-hop scenarios are involved. We also trained question answering and generative commonsense reasoning models using COM2 and found significant improvements in eight diverse downstream tasks. This highlights the potential of using CSKGs to acquire complex reasoning signals inexpensively, without relying on extra human effort.

8.2 Discussions on the Strengths and Limitations

Our proposed commonsense acquisition pipeline has mainly two advantages. First, it is scalable and cheap. The CSKB Population pipeline does not include expensive human annotations or LLM prompting. It builds on existing information-extracted discourse knowledge bases, and uses a classifier fine-tuned on the existing human-annotated CSKBs. It can easily be scaled up. Second, it can provide more diverse and novel knowledge. Our pipeline uses large-scale information-extracted knowledge as candidate knowledge, which is novel and diverse as evaluated in Section 3. The benefits are reflected by the downstream commonsense question answering improvements in Table 6.2.

However, there are also several limitations of CSKB Population. First, with the development of backbone large language models, LLMs can not generate very plausible commonsense knowledge with concrete instructions and several exemplars, and CSKB population cannot outperform LLMs in terms of accuracy. As experiments showed in NovaCOMET and ATOMIC-10x, the annotation quality is even better than humans. Nevertheless, it is not a limitation of our framework but a limitation of the techniques we used. If we apply the latest LLM-powered information extraction tools and the latest LLMs as the backbone classifier, the performance of our framework is also expected to improve a lot. This is left as a future work. Second, the creation of the candidate information-extraction discourse knowledge graph is costly and requires parsing on hundred-billion-scale corpora.

8.3 Future Works

Regarding the acquisition of commonsense knowledge, future works regarding improving commonsense knowledge base population can focus on the denoising of unlabeled discourse knowledge. Even though in PseudoReasoner, we already include some plausibility filter and influence-function based filters, more fine-grained algorithms that focuses on denoising have the potential of further improving the population performance.

In addition, currently, the candidate knowledge come from information extraction, which contain a lot of noise by nature. With the development of backbone language models and their ability of serving as knowledge bases, the process of providing candidate commonsense knowledge can be replaced with prompting large language models.

Regarding reasoning, since our current efforts on COM2 in Chapter 7 only study the benchmarking and plain fine-tuning and in-context learning experiments, future direction can be put to more fine-grained knowledge-augmented methods. For example, the reasoning in COM2 is related to first linking the events or situations in the context to the knowledge base, and the answers may be entailed by the surrounding neighbors in the knowledge base. Future works can be also put to generative retrieval, considering the remarkable effect of language models as knowledge bases and the contextual reasoning ability.

BIBLIOGRAPHY

[1] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran, and J. Chu-Carroll, "Glucose: Generalized and contextualized story explanations," in *The Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020.

[2] C. Malaviya, C. Bhagavatula, A. Bosselut, and Y. Choi, "Commonsense knowledge base completion with structural and semantic context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2925–2933.

[3] P. R. Cohen, *Empirical methods for artificial intelligence*. MIT press Cambridge, MA, 1995, vol. 139.

[4] L. Yao, C. Mao, and Y. Luo, "KG-BERT: BERT for knowledge graph completion," *CoRR*, vol. abs/1909.03193, 2019.

[5] Q. Xie, Z. Dai, E. H. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[6] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. L. Bras, J. Wang, C. Bhagavatula, Y. Choi, and D. Downey, "G-daug: Generative data augmentation for commonsense reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, ser. Findings of ACL, T. Cohn, Y. He, and Y. Liu, Eds., vol. EMNLP 2020. Association for Computational Linguistics, 2020, pp. 1008–1025.

[7] Q. Xie, M. Luong, E. H. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *2020 IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020, pp. 10 684–10 695.

[8] S. Ravi, R. Ng, and V. Shwartz, "COMET-M: reasoning about multiple events in complex sentences," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 12 921–12 937. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.861

[9] S. Gabriel, C. Bhagavatula, V. Shwartz, R. L. Bras, M. Forbes, and Y. Choi, "Paragraph-level commonsense transformers with recurrent memory," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 2021, pp. 12 857–12 865. [Online]. Available: https://doi.org/10.1609/aaai.v35i14.17521

[10] J. D. Hwang, C. Bhagavatula, R. L. Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi, "(comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16792

[11] H. Zhang, X. Liu, H. Pan, H. Ke, J. Ou, T. Fang, and Y. Song, "Aser: Towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities," *arXiv preprint arXiv:2104.02137*, 2021.

[12] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[13] J. McCarthy, "Artificial intelligence, logic and formalizing common sense," *Philosophical Logic and Artificial Intelligence*, pp. 161–190, 1989.

[14] E. Davis, *Representations of commonsense knowledge*, ser. notThenot Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1990.

[15] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *AAAI*, 2018, pp. 4970–4977.

[16] S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu, "Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, 2020, pp. 3766–3772.

[17] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1421

[18] A. Talmor, O. Yoran, R. L. Bras, C. Bhagavatula, Y. Goldberg, Y. Choi, and J. Berant, "Commonsenseqa 2.0: Exposing the limits of AI through gamification," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, J. Vanschoren and S. Yeung, Eds., 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/3ef815416f775098fe977004015c6193-Abstract-round1.html

[19] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: an adversarial winograd schema challenge at scale," *Commun. ACM*, vol. 64, no. 9, 2021. [Online]. Available: https://doi.org/10.1145/3474381

[20] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial

dataset for grounded commonsense inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 93–104.

[21] M. Sap, H. Rashkin, D. Chen, R. L. Bras, and Y. Choi, "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 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1454

[22] B. Zhou, Q. Ning, D. Khashabi, and D. Roth, "Temporal common sense acquisition with minimal supervision," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7579–7589.

[23] X. Liu, D. Yin, Y. Feng, and D. Zhao, "Things not written in text: Exploring spatial commonsense from visual signals," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 2365–2376. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.168

[24] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "AGQA 2.0: An updated benchmark for compositional spatio-temporal reasoning," *CoRR*, vol. abs/2204.06105, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.06105

[25] Y. Elazar, A. Mahabal, D. Ramachandran, T. Bedrax-Weiss, and D. Roth, "How large are lions? inducing distributions over quantitative attributes," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3973–3983.

[26] B. Y. Lin, S. Lee, R. Khanna, and X. Ren, "Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models," in

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 6862–6868. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.557

[27] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "MERLOT: multimodal neural script knowledge models," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 23 634–23 651. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/c6d4eb15f1e84a36eff58eca3627c82e-Abstract.html

[28] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6720–6731.

[29] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017.* IEEE Computer Society, 2017, pp. 5843–5851. [Online]. Available: https://doi.org/10.1109/ICCV.2017.622

[30] H. Kim, A. Zala, and M. Bansal, "Cosim: Commonsense reasoning for counterfactual scene imagination," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 911–923. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.66

[31] D. B. Lenat, “CYC: A large-scale investment in knowledge infrastructure,” *Commun. ACM*, vol. 38, no. 11, pp. 32–38, 1995.

[32] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: an open multilingual graph of general knowledge,” in *AAAI*, 2017, pp. 4444–4451.

[33] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, “Dbpedia: A nucleus for a web of open data,” in *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, ser. Lecture Notes in Computer Science, K. Aberer, K. Choi, N. F. Noy, D. Allemang, K. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, Eds., vol. 4825. Springer, 2007, pp. 722–735. [Online]. Available: https://doi.org/10.1007/978-3-540-76298-0_52

[34] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[35] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[36] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[38] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[39] J. Davison, J. Feldman, and A. M. Rush, "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 2019, Hong Kong, China*, 2019, pp. 1173–1178.

[40] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language models as knowledge bases?" in *EMNLP-IJCNLP*, 2019, pp. 2463–2473.

[41] OpenAI, "Chatgpt: Optimizing language models for dialogue," *OpenAI*, 2022. [Online]. Available: https://openai.com/blog/chatgpt

[42] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.13971

[43] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[44] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[45] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," *CoRR*, vol. abs/2203.15556, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2203.15556

[46] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko,

J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *CoRR*, vol. abs/2204.02311, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.02311

[47] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *CoRR*, vol. abs/2302.04023, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.04023

[48] C. Chan, J. Cheng, W. Wang, Y. Jiang, T. Fang, X. Liu, and Y. Song, "Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations," *CoRR*, vol. abs/2304.14827, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.14827

[49] H. Kim, J. Hessel, L. Jiang, P. West, X. Lu, Y. Yu, P. Zhou, R. L. Bras, M. Alikhani, G. Kim, M. Sap, and Y. Choi, "SODA: million-scale dialogue distillation with social commonsense contextualization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 12 930–12 949. [Online]. Available: https://aclanthology.org/2023.emnlp-main.799

[50] H. He, H. Zhang, and D. Roth, "Rethinking with retrieval: Faithful large language model inference," *CoRR*, vol. abs/2301.00303, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.00303

[51] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *CoRR*, vol. abs/2302.12813, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.12813

[52] J. Gordon and B. V. Durme, "Reporting bias and knowledge acquisition," in *AKBC@CIKM*, 2013, pp. 25–30.

[53] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

[54] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, B. He, S. Jiang, and B. Dong, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," *arXiv preprint arXiv:2303.16421*, 2023.

[55] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *AAAI*, vol. 33, 2019, pp. 3027–3035.

[56] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in *ACL*, 2019, pp. 4762–4779.

[57] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *ICML*, vol. 69, 2004.

[58] H. Zhang, X. Liu, H. Pan, Y. Song, and C. W.-K. Leung, "Aser: A large-scale eventuality knowledge graph," in *WWW*, 2020, pp. 201–211.

[59] H. Zhang, X. Liu, H. Pan, H. Ke, J. Ou, T. Fang, and Y. Song, "Aser: Towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities," *Artificial Intelligence*, vol. 309, p. 103740, 2022.

[60] K. K. Teru, E. Denis, and W. L. Hamilton, "Inductive relation prediction by subgraph reasoning," in *ICML*, 2020.

[61] W. Zhao, M. Geva, B. Y. Lin, M. Yasunaga, A. Madaan, and T. Yu, "Complex reasoning in natural languag," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2023, Toronto, Canada, July 9-14, 2023*, Y. V. Chen, M. Mieskes, and S. Reddy, Eds. Association for Computational Linguistics, 2023, pp. 11–20. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-tutorials.2

[62] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. W. Buchanan, L. Berkowitz, O. Biran, and J. Chu-Carroll, "GLUCOSE: generalized and contextualized story explanations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, 2020, pp. 4569–4586.

[63] W. L. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, and J. Leskovec, "Embedding logical queries on knowledge graphs," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 2030–2041. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/ef50c335cca9f340bde656363ebd02fd-Abstract.html

[64] H. P. Grice, "Logic and conversation," in *Speech acts*. Brill, 1975, pp. 41–58.

[65] D. B. Lenat and R. V. Guha, *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[66] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open Mind Common Sense: Knowledge acquisition from the general public," in *2002 Confederated International Conferences DOA, CoopIS and ODBASE , Irvine, California, USA*, ser. Lecture Notes in Computer Science, vol. 2519. Springer, 2002, pp. 1223–1237.

[67] H. Rashkin, M. Sap, E. Allaway, N. A. Smith, and Y. Choi, "Event2mind: Commonsense inference on events, intents, and reactions," in *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 2018, pp. 463–473.

[68] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, and Y. Choi, "Social chemistry 101: Learning to reason about social and moral norms," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 653–670.

[69] D. Ghosal, S. Shen, N. Majumder, R. Mihalcea, and S. Poria, "CICERO: A dataset for contextualized commonsense inference in dialogues," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022, Dublin, Ireland*, 2022, pp. 5010–5028.

[70] S. Gao, B. Borges, S. Oh, D. Bayazit, S. Kanno, H. Wakaki, Y. Mitsufuji, and A. Bosselut, "Peacok: Persona commonsense knowledge for consistent and engaging narratives," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.362

[71] M. He, T. Fang, W. Wang, and Y. Song, "Acquiring and modelling abstract commonsense knowledge via conceptualization," *arXiv preprint arXiv:2206.01532*, 2022.

[72] Z. Wang, H. Shi, W. Wang, T. Fang, H. Zhang, S. Choi, X. Liu, and Y. Song, "Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph," *CoRR*, vol. abs/2311.09174, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.09174

[73] G. A. Miller, "Wordnet: A lexical database for english," vol. 38, no. 11, 1995, pp. 39–41. [Online]. Available: https://doi.org/10.1145/219717.219748

[74] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding," in *Proceedings of the ACM SIGMOD International Conference on*

Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012, K. S. Candan, Y. Chen, R. T. Snodgrass, L. Gravano, and A. Fuxman, Eds. ACM, 2012, pp. 481–492. [Online]. Available: https://doi.org/10.1145/2213836.2213891

[75] S. Gao, J. D. Hwang, S. Kanno, H. Wakaki, Y. Mitsufuji, and A. Bosselut, "Comfact: A benchmark for linking contextual commonsense knowledge," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 1656–1675.

[76] F. Mu and W. Li, "Enhancing narrative commonsense reasoning with multilevel causal knowledge," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[77] P. Ammanabrolu, W. Cheung, W. Broniec, and M. O. Riedl, "Automated storytelling via causal, commonsense plot ordering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 5859–5867.

[78] B. P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, and J. McAuley, "Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9194–9206.

[79] M. Ismayilzada, D. Paul, S. Montariol, M. Geva, and A. Bosselut, "Crow: Benchmarking commonsense reasoning in real-world tasks," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9785–9821.

[80] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905. Springer, 2016, pp. 852–869. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_51

[81] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual

genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0981-7

[82] J. Hessel, J. D. Hwang, J. S. Park, R. Zellers, C. Bhagavatula, A. Rohrbach, K. Saenko, and Y. Choi, "The abduction of sherlock holmes: A dataset for visual abductive reasoning," in *European Conference on Computer Vision*. Springer, 2022, pp. 558–575.

[83] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in neural information processing systems*, vol. 34, pp. 23 634–23 651, 2021.

[84] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2022, Seattle, WA, USA*, 2022, pp. 5884–5906.

[85] J. Gordon, B. V. Durme, and L. K. Schubert, "Learning from the Web: Extracting general world knowledge from noisy text," in *Collaboratively-Built Knowledge Sources and Artificial Intelligence, Papers from the 2010 AAAI Workshop, Atlanta, Georgia, USA*, ser. AAAI Technical Report, vol. WS-10-02. AAAI, 2010.

[86] N. Tandon, G. de Melo, F. M. Suchanek, and G. Weikum, "WebChild: Harvesting and organizing commonsense knowledge from the web," in *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, USA*. ACM, 2014, pp. 523–532.

[87] N. Tandon, G. de Melo, A. De, and G. Weikum, "Knowlywood: Mining activity knowledge from hollywood narratives," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, Australia*. ACM, 2015, pp. 223–232.

[88] R. Speer, J. Chin, and C. Havasi, “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*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[89] J. Liu, T. Chen, C. Wang, J. Liang, L. Chen, Y. Xiao, Y. Chen, and K. Jin, “Vocsk: Verb-oriented commonsense knowledge mining with taxonomy-guided induction,” *Artificial Intelligence*, vol. 310, p. 103744, 2022.

[90] H. Zhang, X. Liu, H. Pan, Y. Song, and C. W. Leung, “ASER: A large-scale eventuality knowledge graph,” in *WWW ’20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Y. Huang, I. King, T. Liu, and M. van Steen, Eds. ACM / IW3C2, 2020, pp. 201–211.

[91] H. Zhang, X. Liu, H. Pan, H. Ke, J. Ou, T. Fang, and Y. Song, “ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities,” *Artif. Intell.*, vol. 309, 2022. [Online]. Available: https://doi.org/10.1016/j.artint.2022.103740

[92] T.-P. Nguyen, S. Razniewski, and G. Weikum, “Advanced semantics for commonsense knowledge extraction,” in *Proceedings of the Web Conference 2021*, 2021, pp. 2636–2647.

[93] T.-P. Nguyen, S. Razniewski, J. Romero, and G. Weikum, “Refined commonsense knowledge from large-scale web contents,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8431–8447, 2022.

[94] H. Shi and P. Wolff, “What transformers might know about the physical world: T5 and the origins of knowledge,” in *Proceedings of the 43th Annual Meeting of the Cognitive Science Society, CogSci 2021, virtual*, 2021.

[95] N. Weir, A. Poliak, and B. V. Durme, “Probing neural language models for human tacit assumptions,” in *Proceedings of the 42th Annual Meeting of the Cognitive Science Society, CogSci 2020, virtual*, 2020.

[96] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller, “Language models as knowledge bases?” 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*, 2019, pp. 2463–2473.

[97] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, “How can we know what language models know,” *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 423–438, 2020.

[98] T. Shin, Y. Razeghi, R. L. L. IV, E. Wallace, and S. Singh, “AutoPrompt: Eliciting knowledge from language models with automatically generated prompts,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, 2020, pp. 4222–4235.

[99] Z. Zhong, D. Friedman, and D. Chen, “Factual probing is [MASK]: learning vs. learning to recall,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online*, 2021, pp. 5017–5033.

[100] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “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*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[101] OpenAI, “GPT-4 technical report,” *CoRR*, vol. abs/2303.08774, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.08774

[102] P. West, C. Bhagavatula, J. Hessel, J. D. Hwang, L. Jiang, R. L. Bras, X. Lu, S. Welleck, and Y. Choi, “Symbolic knowledge distillation: from general language models to commonsense models,” in *Proceedings of the 2022*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 4602–4625. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.341

[103] P. West, R. Bras, T. Sorensen, B. Lin, L. Jiang, X. Lu, K. Chandu, J. Hessel, A. Baheti, C. Bhagavatula, and Y. Choi, "NovaCOMET: Open commonsense foundation models with symbolic knowledge distillation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1127–1149. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.80

[104] A. Tamborrino, N. Pellicanò, B. Pannier, P. Voitot, and L. Naudin, "Pre-training is (almost) all you need: An application to commonsense reasoning," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*, 2020, pp. 3878–3887.

[105] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[106] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Neural Information Processing Systems (NIPS)*, 2013, pp. 1–9.

[107] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[108] Z. Sun, Z. Deng, J. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *7th International Conference on Learning Rep-*

resentations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[109] C. Shang, Y. Tang, J. Huang, J. Bi, X. He, and B. Zhou, "End-to-end structure-aware convolutional networks for knowledge base completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3060–3067.

[110] T. Fang, H. Zhang, W. Wang, Y. Song, and B. He, "DISCOS: bridging the gap between discourse knowledge and commonsense knowledge," in *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and L. Zia, Eds. ACM / IW3C2, 2021, pp. 2648–2659.

[111] L. Yao, C. Mao, and Y. Luo, "Kg-bert: Bert for knowledge graph completion," *arXiv preprint arXiv:1909.03193*, 2019.

[112] X. Li, A. Taheri, L. Tu, and K. Gimpel, "Commonsense knowledge base completion," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. [Online]. Available: https://doi.org/10.18653/v1/p16-1137

[113] I. Saito, K. Nishida, H. Asano, and J. Tomita, "Commonsense knowledge base completion and generation," in *CONLL*, 2018, pp. 141–150.

[114] B. Wang, G. Wang, J. Huang, J. You, J. Leskovec, and C.-C. J. Kuo, "Inductive learning on commonsense knowledge graph completion," *arXiv preprint arXiv:2009.09263*, 2020.

[115] T. Fang, W. Wang, S. Choi, S. Hao, H. Zhang, Y. Song, and B. He, "Benchmarking commonsense knowledge base population with an effective evaluation dataset," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih,

Eds. Association for Computational Linguistics, 2021, pp. 8949–8964. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.705

[116] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017, pp. 1024–1034.

[117] L. Yao, C. Mao, and Y. Luo, "KG-BERT: BERT for knowledge graph completion," *CoRR*, vol. abs/1909.03193, 2019. [Online]. Available: http://arxiv.org/abs/1909.03193

[118] T. Fang, Q. V. Do, H. Zhang, Y. Song, G. Y. Wong, and S. See, "Pseudoreasoner: Leveraging pseudo labels for commonsense knowledge base population," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 3379–3394. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.246

[119] M. Roemmele, C. A. Bejan, and A. S. Gordon, "Choice of plausible alternatives: An evaluation of commonsense causal reasoning." in *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011, pp. 90–95.

[120] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge targeting commonsense knowledge," in *NAACL-HLT*, 2019, pp. 4149–4158.

[121] L. Huang, R. Le Bras, C. Bhagavatula, and Y. Choi, "Cosmos qa: Machine reading comprehension with contextual 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, pp. 2391–2401.

[122] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie, "Dream: A challenge data set and models for dialogue-based reading comprehension," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 217–231, 2019.

[123] M. Forbes and Y. Choi, “Verb physics: Relative physical knowledge of actions and objects,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 266–276.

[124] B. Y. Lin, S. Lee, R. Khanna, and X. Ren, “Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6862–6868.

[125] Y. Bisk, R. Zellers, R. LeBras, J. Gao, and Y. Choi, “PIQA: reasoning about physical commonsense in natural language,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020.* AAAI Press, 2020, pp. 7432–7439.

[126] Z. Sun, Z. Deng, J. Nie, and J. Tang, “Rotate: Knowledge graph embedding by relational rotation in complex space,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=HkgEQnRqYQ

[127] Q. Lin, R. Mao, J. Liu, F. Xu, and E. Cambria, “Fusing topology contexts and logical rules in language models for knowledge graph completion,” *Inf. Fusion*, vol. 90, pp. 253–264, 2023. [Online]. Available: https://doi.org/10.1016/j.inffus.2022.09.020

[128] H. Ren, W. Hu, and J. Leskovec, “Query2box: Reasoning over knowledge graphs in vector space using box embeddings,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=BJgr4kSFDS

[129] Z. Wang, H. Yin, and Y. Song, “Benchmarking the combinatorial generalizability of complex query answering on knowledge graphs,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, J. Vanschoren and S. Yeung,

Eds., 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/7eabe3a1649ffa2b3ff8c02ebfd5659f-Abstract-round2.html

[130] Z. Wang, Y. Song, G. Y. Wong, and S. See, "Logical message passing networks with one-hop inference on atomic formulas," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=SoyOsp7i_l

[131] J. Bai, X. Liu, W. Wang, C. Luo, and Y. Song, "Complex query answering on eventuality knowledge graph with implicit logical constraints," *CoRR*, vol. abs/2305.19068, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.19068

[132] H. Ren and J. Leskovec, "Beta embeddings for multi-hop logical reasoning in knowledge graphs," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/e43739bba7cdb577e9e3e4e42447f5a5-Abstract.html

[133] J. Bai, Z. Wang, H. Zhang, and Y. Song, "Query2particles: Knowledge graph reasoning with particle embeddings," in *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 2703–2714. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-naacl.207

[134] Y. Bai, X. Lv, J. Li, and L. Hou, "Answering complex logical queries on knowledge graphs via query computation tree optimization," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho,

B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 1472–1491. [Online]. Available: https://proceedings.mlr.press/v202/bai23b.html

[135] X. Guan, B. Cao, Q. Gao, Z. Yin, B. Liu, and J. Cao, "Multi-hop commonsense knowledge injection framework for zero-shot commonsense question answering," *CoRR*, vol. abs/2305.05936, 2023. [Online]. Available: http://arxiv.org/abs/2305.05936

[136] W. Ding, S. Feng, Y. Liu, Z. Tan, V. Balachandran, T. He, and Y. Tsvetkov, "Knowledge crosswords: Geometric reasoning over structured knowledge with large language models," *arXiv preprint arXiv:2310.01290*, 2023.

[137] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J. Wen, "Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph," *CoRR*, vol. abs/2402.11163, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.11163

[138] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, "Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 346–361, 2021.

[139] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Musique: Multihop questions via single-hop question composition," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 539–554, 2022.

[140] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2369–2380.

[141] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[142] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic, "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026.

[143] B. Y. Lin, X. Chen, J. Chen, and X. Ren, "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, pp. 2829–2839.

[144] Y. Feng, X. Chen, B. Y. Lin, P. Wang, J. Yan, and X. Ren, "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, pp. 1295–1309.

[145] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "Qa-gnn: Reasoning with language models and knowledge graphs for question answering," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

[146] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec, "Greaselm: Graph reasoning enhanced language models," in *International Conference on Learning Representations*, 2021.

[147] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[148] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[149] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing*

Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

[150] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le *et al.*, "Least-to-most prompting enables complex reasoning in large language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[151] S. Diao, P. Wang, Y. Lin, and T. Zhang, "Active prompting with chain-of-thought for large language models," *arXiv preprint arXiv:2302.12246*, 2023.

[152] Z. Ling, Y. Fang, X. Li, Z. Huang, M. Lee, R. Memisevic, and H. Su, "Deductive verification of chain-of-thought reasoning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[153] H. S. Zheng, S. Mishra, X. Chen, H.-T. Cheng, E. H. Chi, Q. V. Le, and D. Zhou, "Take a step back: Evoking reasoning via abstraction in large language models," *arXiv preprint arXiv:2310.06117*, 2023.

[154] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[155] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *NAACL-HLT*, Jun. 2016, pp. 110–119.

[156] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, "Generating informative and diverse conversational responses via adversarial information maximization," in *NeurIPS*, 2018, pp. 1810–1820.

[157] F. Moghimifar, L. Qu, Y. Zhuo, G. Haffari, and M. Baktashmotlagh, "Neural-symbolic commonsense reasoner with relation predictors," *arXiv preprint arXiv:2105.06717*, 2021.

[158] N. Tandon, G. De Melo, A. De, and G. Weikum, "Knowlywood: Mining activity knowledge from hollywood narratives," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 223–232.

[159] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[160] E. Davis, "Benchmarks for automated commonsense reasoning: A survey," *CoRR*, vol. abs/2302.04752, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.04752

[161] J. Liu, W. Wang, D. Wang, N. A. Smith, Y. Choi, and H. Hajishirzi, "Vera: A general-purpose plausibility estimation model for commonsense statements," *arXiv preprint arXiv:2305.03695*, 2023.

[162] K. Ma, F. Ilievski, J. Francis, Y. Bisk, E. Nyberg, and A. Oltramari, "Knowledge-driven data construction for zero-shot evaluation in commonsense question answering," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17593

[163] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[164] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite,

J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, Q. Liu and D. Schlangen, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-demos.6

[165] X. Li, A. Taheri, L. Tu, and K. Gimpel, "Commonsense knowledge base completion," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[166] I. Saito, K. Nishida, H. Asano, and J. Tomita, "Commonsense knowledge base completion and generation," in *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, A. Korhonen and I. Titov, Eds. Association for Computational Linguistics, 2018, pp. 141–150.

[167] C. Malaviya, C. Bhagavatula, A. Bosselut, and Y. Choi, "Commonsense knowledge base completion with structural and semantic context," 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*. AAAI Press, 2020, pp. 2925–2933.

[168] T. Fang, W. Wang, S. Choi, S. Hao, H. Zhang, Y. Song, and B. He, "Benchmarking commonsense knowledge base population with an effective evaluation dataset," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.705

[169] V. Shwartz, P. West, R. L. Bras, C. Bhagavatula, and Y. Choi, "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*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 4615–4629.

[170] A. Bosselut, R. L. Bras, and Y. Choi, "Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 2021, pp. 4923–4931.

[171] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 1885–1894.

[172] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," 2019.

[173] X. Han, B. C. Wallace, and Y. Tsvetkov, "Explaining black box predictions and unveiling data artifacts through influence functions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 5553–5563.

[174] N. Agarwal, B. Bullins, and E. Hazan, "Second-order stochastic optimization for machine learning in linear time," *J. Mach. Learn. Res.*, vol. 18, pp. 116:1–116:40, 2017.

[175] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettle-

moyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019.

[176] P. He, X. Liu, J. Gao, and W. Chen, “Deberta: decoding-enhanced bert with disentangled attention,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[177] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7871–7880.

[178] P. West, C. Bhagavatula, J. Hessel, J. D. Hwang, L. Jiang, R. L. Bras, X. Lu, S. Welleck, and Y. Choi, “Symbolic knowledge distillation: from general language models to commonsense models,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 4602–4625.

[179] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[180] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[181] A. Lavie and A. Agarwal, “METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proceedings of the Second*

Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007, C. Callison-Burch, P. Koehn, C. S. Fordyce, and C. Monz, Eds. Association for Computational Linguistics, 2007, pp. 228–231. [Online]. Available: https://aclanthology.org/W07-0734/

[182] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 4566–4575. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7299087

[183] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. L. Bras, J. Wang, C. Bhagavatula, Y. Choi, and D. Downey, “G-daug: Generative data augmentation for commonsense reasoning,” in *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, ser. Findings of ACL, T. Cohn, Y. He, and Y. Liu, Eds., vol. EMNLP 2020. Association for Computational Linguistics, 2020, pp. 1008–1025. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.90

[184] X. Peng, S. Li, S. Wiegreffe, and M. Riedl, “Inferring the reader: Guiding automated story generation with commonsense reasoning,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7008–7029. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.520

[185] Q. Tu, Y. Li, J. Cui, B. Wang, J.-R. Wen, and R. Yan, “MISC: A mixed strategy-aware model integrating COMET for emotional support conversation,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 308–319. [Online]. Available: https://aclanthology.org/2022.acl-long.25

[186] P. He, J. Gao, and W. Chen, “DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing,” in *The Eleventh*

International Conference on Learning Representations, 2023. [Online]. Available: https://openreview.net/forum?id=sE7-XhLxHA

[187] P. Banerjee and C. Baral, "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*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.11

[188] Y. Su, Z. Wang, T. Fang, H. Zhang, Y. Song, and T. Zhang, "MICO: A multi-alternative contrastive learning framework for commonsense knowledge representation," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.96

[189] Y. J. Kim, B. Kwak, Y. Kim, R. K. Amplayo, S. Hwang, and J. Yeo, "Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.163

[190] C. Bhagavatula, R. L. Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W. Yih, and Y. Choi, "Abductive commonsense reasoning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=Byg1v1HKDB

[191] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "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.* AAAI Press, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33013027

[192] C. Malaviya, C. Bhagavatula, A. Bosselut, and Y. Choi, "Commonsense knowledge base completion with structural and semantic context," 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.* AAAI Press, 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5684

[193] J. Liu, W. Wang, D. Wang, N. A. Smith, Y. Choi, and H. Hajishirzi, "Vera: A general-purpose plausibility estimation model for commonsense statements," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 1264–1287. [Online]. Available: https://aclanthology.org/2023.emnlp-main.81

[194] W. Wang, T. Fang, W. Ding, B. Xu, X. Liu, Y. Song, and A. Bosselut, "CAR: Conceptualization-augmented reasoner for zero-shot commonsense question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 520–13 545. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.902

[195] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami,

N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *CoRR*, vol. abs/2307.09288, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.09288

[196] D. Khashabi, Y. Kordi, and H. Hajishirzi, "Unifiedqa-v2: Stronger generalization via broader cross-format training," *CoRR*, vol. abs/2202.12359, 2022. [Online]. Available: https://arxiv.org/abs/2202.12359

[197] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *CoRR*, vol. abs/2210.11416, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.11416

[198] J. Robinson, C. M. Rytting, and D. Wingate, "Leveraging large language models for multiple choice question answering," *CoRR*, vol. abs/2210.12353, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.12353

[199] Z. Chen, A. Hernández-Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M. Hartley, M. Jaggi, and A. Bosselut, "MEDITRON-70B: scaling medical pretraining for large language models," *CoRR*, vol. abs/2311.16079, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.16079

APPENDIX A

ADDITIONAL DETAILS ON COMMONSENSE KNOWLEDGE BASES

A.1 Relations

During human annotation, we translate the symbolic knowledge triples into human language for annotators to better understand the questions. A (h, r, t) triple where h, r, and t are the head, relation, and tail, is translated to *if* h*, then [Description],* t. Here, the description placeholder *[Description]* comes from rules in Table A.1, which is modified from [10]. These descriptions can also be regarded as definitions of those commonsense relations.

Moreover, the definitions of the discourse relations in ASER are presented in Table A.2. We also present the statistics of relation distribution for ASER$_{norm}$ in Table A.3.

A.2 Additional Details of Pre-processing

A.2.1 Examples of Format Unification

Table A.4 demonstrates several examples for unifying the formats of different resources. In ConceptNet and Knowlywood, the nodes are mostly *verb* or *verb-object* phrases, and we add a subject "*PersonX*" in front of each node. For ATOMIC, the main modification part is the tails, where subjects tend to be missing. We treat *agent*-driven (relations investigating causes and effects on *PersonX*) and *theme*-driven (relations investigating causes and effects on *PersonY*) differently, and add *PersonX* or *PersonY* in front of the tails whose subjects are missing. For ASER, rules are used to discriminate *PersonX* and *PersonY* in a certain edge. Two examples for ASER and ATOMIC demonstrating the differences between *PersonX* and *PersonY* are provided in the table. For GLUCOSE, we simply replace *SomeoneA* with

Relation	Decriptions
oEffect	then, *PersonY* will
xEffect	then, *PersonX* will
gEffect	then, other people or things will
oWant	then, *PersonY* wants to
xWant	then, *PersonX* wants to
gWant	then, other people or things want to
oReact	then, *PersonY* feels
xReact	then, *PersonX* feels
gReact	then, other people or things feel
xAttr	*PersonX* is seen as
xNeed	but before, *PersonX* needed
xIntent	because *PersonX* wanted
isBefore	happens before
isAfter	happens after
HinderedBy	can be hindered by
xReason	because
Causes	causes
HasSubEvent	includes the event/action

Table A.1: Descriptions of different commonsense relations, which are translation rules from knowledge triples (h, r, t) to human language, "*if h, then [Description], t*" [10].

PersonX and *SomeoneB* with *PersonY* accordingly. Moreover, all the words are lemmatized using Stanford CoreNLP parser[1] to normalized forms.

A.2.2 Examples of Populated Triples

Examples of the annotations of the populated triples are listed in Table A.7. In the *Original Test Set* category, the triples are composed of two parts, one is the ground truth triples from the original CSKBs, and one is triples randomly sampled from $\mathcal{G}^c$.

[1] https://stanfordnlp.github.io/CoreNLP/

Relation	Decriptions
`Precedence`	h happens before t
`Succession`	h happens after h
`Synchronous`	h happens the same time as t
`Reason`	h happens because t
`Result`	h result in t
`Condition`	Only when t happens, h can happen
`Contrast`	h and t share significant difference regarding some property
`Concession`	h and t result in another opposite event
`Alternative`	h and t are alternative situations of each other.
`Conjunction`	h and t both happen
`Restatement`	h restates t
`Instantiation`	t is a more detailed description of h
`ChosenAlternative`	h and t are alternative situations of each other, but the subject prefers h
`Exception`	t is an exception of h
`Co_Occurrence`	h and t co-occur at the same sentence

Table A.2: Descriptions of discourse relations in ASER [11].

Relation	number of edges
`Precedence`	4,957,481
`Succession`	1,783,154
`Synchronous`	8,317,572
`Reason`	5,888,968
`Result`	5,562,565
`Condition`	8,109,020
`Contrast`	23,208,195
`Concession`	1,189,167
`Alternative`	1,508,729
`Conjunction`	37,802,734
`Restatement`	159,667
`Instantiation`	33,840
`ChosenAlternative`	91,286
`Exception`	51,502
`Co_Occurrence`	124,330,714
Total	222,994,594

Table A.3: Statistics of relations in ASER$_{norm}$.

Resource	Original Format			Aligned Format	
	Head	Relation	Tail	Head	Tail
ConceptNet	get exercise	`HasSubEvent`	ride bicycle	*PersonX* get exercise	*PersonX* ride bicycle
ATOMIC$\binom{20}{20}$	*PersonX* gets exercise	`xReact`	tired	*PersonX* get exercise	*PersonX* be tired
	PersonX visits *PersonY* at work	`oEffect`	say hello	*PersonX* visits *PersonY*	*PersonY* say hello
GLUCOSE	*SomeoneA* gets exercise	`Dim 1 (xEffect)`	*SomeoneA* gets tired	*PersonX* get exercise	*PersonX* be tired
Knowlywood	get exercise	`NextActivity`	take shower	*PersonX* get exercise	*PersonX* take shower
ASER	he gets exercise	`Result`	he is tired	*PersonX* get exercise	*PersonX* be tired
	he visits her at work	`Precedence`	she is happy	*PersonX* visit *PersonY* at work	*PersonY* is happy

Table A.4: Examples of format unification of CSKBs and eventuality graphs.

Commonsense Relations	ASER Relations	Patterns
Effect, Want isBefore, Causes	Result, Precedence, Condition^{-1}, Succession^{-1}, Reason^{-1}	-
React	Result, Precedence, Condition^{-1}, Succession^{-1}, Reason^{-1}	*s-v/be-a/o, s-v-be-a/o, s-v, spass-v*
xIntent, xNeed, isAfter	Condition, Succession, Reason, Result^{-1}, Precedence^{-1}	-
xAttr	Synchronous$^{\pm1}$, Reason$^{\pm1}$, Result$^{\pm1}$, Condition$^{\pm1}$, Conjunction$^{\pm1}$, Restatement$^{\pm1}$	*s-be-a/o, s-v-a, s-v-be-a/o, s-v, spass-v*
HinderedBy	Concession, Alternative	-
HasSubEvent	Synchronous$^{\pm1}$, Conjunction$^{\pm1}$	-

Table A.5: Rules of selecting candidate triples. For a certain commonsense relation $\mathtt{r}_{cs}$ in the first column, (*head*, $\mathtt{r}_{ASER}$, *tail*) in ASER, where $\mathtt{r}_{ASER}$ belongs to the corresponding cell in the second column, can be selected as a candidate (*head*, $\mathtt{r}_{cs}$, *tail*) for annotation.

Model	Average AUC
KG-BERTSAGE (Dir)	66.2
KG-BERTSAGE (Undir)	**67.2**

Table A.6: Experimental results using two different neighboring functions.

Head	Relation	Tail	Label	Source
PersonX give *PersonY* ride	xNeed	*PersonX* need to wear proper clothes	Plau.	Triples in CSKBs (*Original Test Set*)
PersonX be wait for taxi	isAfter	*PersonX* hail a taxi	Plau.	
PersonX be diagnose with something	Causes	*PersonX* be sad	Plau.	
PersonX feel something	xEffect	*PersonX* figure	Implau.	Randomly sampled examples
PersonX be patient with ignorance	HinderedBy	*PersonY* have the right vocabulary	Implau.	
PersonY grasp *PersonY* meaning	HasSubEvent	*PersonY* open it mechanically	Implau.	
PersonX spill coffee	oEffect	*PersonY* have to server	Plau.	*CSKB head + ASER tail*
PersonX care for *PersonY*	xNeed	*PersonX* want to stay together	Plau.	
PersonX be save money	HasSubEvent	PeopleX can not afford something	Plau.	
PersonX decide to order a pizza	xReact	*PersonX* have just move	Implau.	
it be almost christmas	gReact	*PersonX* be panic	Implau.	
arm be break	isBefore	*PersonY* ask	Implau.	
PersonX go early in morning	xEffect	*PersonX* do not have to deal with crowd	Plau.	*ASER edges*
PersonX have take time to think it over *PersonX*	xReact	*PersonX* be glad	Plau.	
PersonX have a good work-life balance	xIntent	*PersonX* be happy	Plau.	
PersonX weight it by value	oWant	*PersonY* bet	Implau.	
PersonX be hang out on reddit	oReact	*PersonY* can not imagine	Implau.	
PersonX can get *PersonY* out shell	xIntent	*PersonX* just start poach *PersonY*	Implau.	

Table A.7: Examples of the human-annotated populated triples.