Social Commonsense Reasoning with Multi-Head Knowledge Attention

Debjit Paul
Research Training Group AIPHES
Institute for Computational Linguistics
Heidelberg University
paul@cl.uni-heidelberg.de

Anette Frank
Research Training Group AIPHES
Institute for Computational Linguistics
Heidelberg University
frank@cl.uni-heidelberg.de

Abstract

Social Commonsense Reasoning requires understanding of text, knowledge about social events and their pragmatic implications, as well as commonsense reasoning skills. In this work we propose a novel *multi-head knowledge attention model* that encodes semi-structured commonsense inference rules and learns to incorporate them in a transformer-based reasoning cell. We assess the model's performance on two tasks that require different reasoning skills: *Abductive Natural Language Inference* and *Counterfactual Invariance Prediction* as a new task. We show that our proposed model improves performance over strong state-of-the-art models (i.e., RoBERTa) across both reasoning tasks. Notably we are, to the best of our knowledge, the first to demonstrate that a model that learns to perform *counterfactual reasoning* helps predicting the best explanation in an *abductive reasoning* task. We validate the robustness of the model's reasoning capabilities by perturbing the knowledge and provide qualitative analysis on the model's knowledge incorporation capabilities.

1 Introduction

Humans are able to understand natural language text about everyday situations effortlessly, by relying on commonsense knowledge and making inferences. For example in Figure 1, given two observations: *Dotty was being very grumpy* and *She felt much better afterwards* – we can come up with a plausible explanation about what could have provoked the change in Dotty's emotion. We can also construct alternative hypotheses that will not change Dotty's emotion. In order to judge the plausibility of such explanations, we need to have information about mental states and social norms, i.e., a form of commonsense knowledge. Such information includes that *calling a close friend*, in

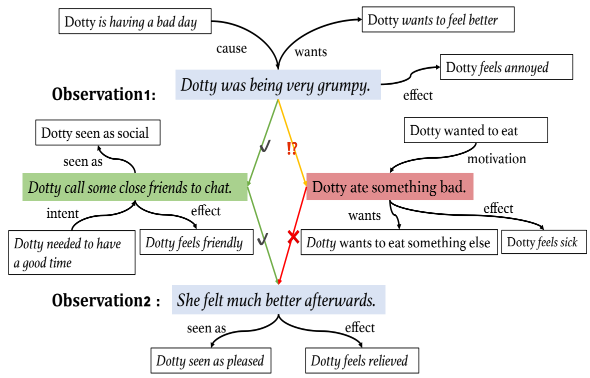

Figure 1: Motivational example: The top and bottom blue boxes show two observations. The green and red box contain a plausible and an implausible hypothesis, respectively. A green line denotes that an event is likely to follow, the yellow line that an event is somewhat unlikely to follow, the red line something unlikely.

general, makes *people feel happy*. This kind of inference goes beyond the broadly studied textual entailment task (Bowman et al., 2015) in that i) it requires a specific form of knowledge, namely knowledge about mental states (*intent, motivation*), social norms (*cause or effect of an event*) and behaviour (*emotional reactions*), and ii) the awareness that inferences we can draw on their basis must often be viewed as plausible explanations, and hence can be defeasible, rather than being strict inferences.

In this paper, we investigate social commonsense reasoning in narrative contexts. Specifically, we address two different reasoning tasks: language-based abductive reasoning, and counterfactual invariance prediction. We introduce the Counterfactual Invariance Prediction task (CIP), which tests the capability of models to predict whether under the assumption of a counterfactual event, a factual event remains invariant or not in a narrative context. Figure 1 illustrates an example: Given a narrative context – *"Dotty was being very grumpy" (premise), "Dotty called some close friends to chat" (hypothe-*

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2969–2980
November 16 - 20, 2020.

Task	Context	Answer
αNLI	O_1: *Dotty was being very grumpy.* H_1: Dotty ate something bad. H_2: Dotty call some close friends to chat. O_2: *She felt much better afterwards.*	H_1 or $\mathbf{H_2}$
CIP	s_1*: Bob had to get to work in the morning.* s_2*: His car battery was struggling to start the car.* s_3*: He called his neighbor for a jump start.* s'_2*: His car won't start.* s_3*: He called his neighbor for a jump start.*	**[Yes]** or [No]
	s_1*: Bill and Teddy were at the bar together.* s_2*: Bill noticed a pretty girl.* s_3*: He went up to her to flirt.* s'_2*: Bill noticed his mom was there.* s_3*: He went up to her to flirt.*	[Yes] or **[No]**

Table 1: Examples from each dataset used in this work. The correct choice in each example is given in bold text.

sis), "She felt much better afterwards."(conclusion) – will a counterfactual assumption (alternative hypothesis), e.g., *"Dotty ate something bad"*, still lead to same conclusion?

While there has been positive impact of transformer-based pretrained language models (LMs) (Devlin et al., 2019; Liu et al., 2019) on several downstream NLU tasks including commonsense reasoning, there is still a performance gap between machines and humans, especially when the task involves implicit knowledge (Talmor et al., 2018).

There are two important bottlenecks: (i) obtaining relevant commonsense knowledge and (ii) effectively incorporating it into state-of-the-art neural models to improve their reasoning capabilities. In current research, the standard approach to address the first bottleneck is to extract knowledge tuples or paths from large structured knowledge graphs (KGs) (e.g. ConceptNet, Speer et al. (2017)) using graph-based methods (Bauer et al., 2018; Paul and Frank, 2019; Lin et al., 2019). However, in this work, instead of retrieving and selecting knowledge from a static KG, we *dynamically generate contextually relevant knowledge* using COMET (based on GPT-2) (Bosselut et al., 2019). To address the second bottleneck, we build on the hypothesis that models performing such reasoning tasks need to consider multiple knowledge rules jointly (see Fig. 1). Hence, we introduce a novel *multi-head knowledge attention model* which learns to focus on multiple pieces of knowledge at the same time, and is able to refine the input representation in a recursive manner, to improve the reasoning capabilities.

An important aspect of using specified knowledge rules is a gain in interpretability. In this work, we perturb the pieces of knowledge available to the model to demonstrate its robustness, and we provide qualitative analysis to offer deeper insight into the model's capabilities.

Our contributions are: i) We propose a new *multi-head knowledge attention model* that uses structured knowledge rules to emulate reasoning. ii) We compare our model with several state-of-the-art neural architectures for QA tasks and show that it performs better on *two types of reasoning tasks*. iii) We specifically compare our novel *knowledge integration technique* to prior integration methods and show it performs better on the abductive reasoning task (+2 percentage points). iv) We introduce a *novel counterfactual invariance prediction* (CIP) task, and show a *correlation* between *abduction* and *counterfactual reasoning* in a narrative context. v) To *analyze the reasoning capabilities* of our model we investigate a) how it performs without fine-tuning on a pre-trained model, b) how robustly it behaves when confronted with perturbations and noise in the knowledge and c) offer qualitative analysis of the reasoning module.

Our code is made publicly available.[1]

2 Social Commonsense Reasoning Tasks

We address two social commonsense reasoning tasks that require different reasoning skills. They are exemplified in Table 1 and detailed below.

Abdutive Natural Language Inference (αNLI) Bhagavatula et al. (2020) created a dataset that tests a model's ability to choose the best explanation for an incomplete set of observations. Abduction is a backward reasoning task. Given a pair of observations O_1 and O_2, the αNLI task is to select the most plausible explanation (hypothesis) H_1 or H_2.

Counterfactual Invariance Prediction (CIP) Counterfactual Reasoning (CR) is the mental abil-

[1] https://github.com/Heidelberg-NLP/MHKA

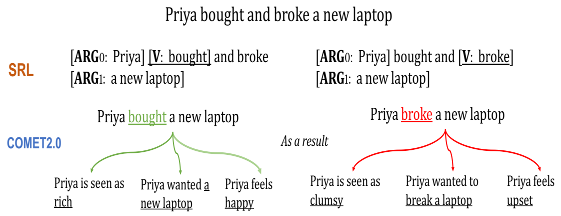

Figure 2: Depicting the steps to extract commonsense knowledge about social events.

ity to construct alternatives (i.e., counterfactual assumptions) to past events and to reason about their (hypothetical) implications (Epstude and Roese, 2008; Roese and Morrison, 2009). One of the key challenges of CR is judging *causal invariance*, i.e., deciding whether a given factual event is invariant under counterfactual assumptions, or whether it is not (Peters et al., 2016; Qin et al., 2019).

In this work, we define a new *Counterfactual Invariance Prediction (CIP)* task that tests the capability of models to predict whether under the assumption of a counterfactual event, a (later) factual event remains invariant or not in a narrative context (cf. Table 1). This task requires deeper understanding of causal narrative chains and reasoning in forward direction. Qin et al. (2019) proposed a dataset to encourage models to learn to rewrite stories with counterfactual reasoning. We automatically collect counterfactual invariance examples along with non-invariant examples from their dataset to create a balanced dataset for our proposed CIP task.

The formal setup is: given the first three consecutive sentences from a narrative story s_1 (premise), s_2 (initial context), s_3 (factual event) and an additional sentence $s_2^{'}$ that is counterfactual to the initial context s_2, the task is to predict whether s_3 is invariant given s_1, $s_2^{'}$ or not. The train/dev/test data (cf. Table 3) are balanced with an equal number of *Yes/No* answers, hence the random baseline is 50%. To compute human performance, we gave 100 instances from the test set to expert evaluators. Human accuracy on the CIP task is at *84.8%*.[2]

3 Semantic & Commonsense Knowledge

This section details the steps we follow to generate social commonsense knowledge about events mentioned in a narrative. See Figure 2 for illustration.

Understanding a narrative text requires the ability to identify events and to reason about their causal effects. Beyond causal relations, they require the understanding of narrative relations, as in narrative chains or schemata (Chambers and Jurafsky, 2008). This is knowledge about characteristic script-like event sequences where semantic roles of consecutive events are referentially bound to roles of preceding events. While Chambers and Jurafsky (2008) focused on the induction of schemata using corpus statistics, we will combine detected events with deeper commonsense knowledge.

In a first step we apply SRL to extract the basic structure *"who did what to whom, when and where"* from each sentence in the context, using state-of-the-art SRL (Shi and Lin, 2019). In a second step, we use commonsense transformer (COMET2.0,[3] Bosselut et al. (2019)) to extract social commonsense knowledge about the extracted events. COMET2.0 is trained on the ATOMIC (Sap et al., 2019) inferential knowledge resource which consists of 877K everyday events, each characterized by nine relation types (*xIntent, xNeed, xReact*, etc.) which we call *dimensions*. These dimensions connect the event in question with manifold properties, emotions, as well as other states or events.

In the last processing step we generate, for each event in each sentence from our datasets, all dimensions defined for it using COMET2.0. For example, for: *Dotty ate something bad* we generate (among others)[4] the tuple: ⟨*PersonX, xReact, sick*⟩ and derive ⟨*Dotty, feels, sick*⟩ by substituting *PersonX* with the logical subject, the filler of the role *ARG0*.

4 A Multi-Head Knowledge Attention (MHKA) Model for Social Reasoning

In this section we introduce the MHKA model and discuss some key differences in how MHKA works for the two different Social Commonsense Reasoning tasks. For a model overview see Figure 3.

4.1 Model Architecture

MHKA consists of 3 modules: (a) the *Context Encoding Layer* consists of a pre-trained LM, (b) the *Knowledge Encoding Layer* consists of stacked transformer blocks, (c) the *Reasoning Cell* consists of transformer blocks with *multi-head attention* that allows the model to jointly attend to the input representation and the encoded knowledge. The input format for each task is depicted in Table 2.

[2] More details about the data are given in the *Supplement*.

[3] COMET2.0 uses GPT-2 as pretrained model.

[4] More examples are given in the *Supplement*.

Task	Input Format	Output
αNLI	[CLS] O_1 H_i [SEP] O_2 [SEP]	H_1 or H_2
CIP	[CLS] s_1 s_2 s_3 [SEP] s_1 s_2' s_3 [SEP]	YES or NO

Table 2: Different input and output formats: [CLS] is a special token used for classification, [SEP] a delimiter.

(a) Context Encoding Layer: For each task, we concatenate the inputs as a sequence of tokens x_n = (x_{n_1}, .. x_{n_m}), and compute contextualized representations with a pre-trained LM. We obtain n different representations for n input options i.e., $h_{x_n} = encode(x_n) = (h_{n_1}, .., h_{n_m})$, where for αNLI n=2 and for CIP n=1. As pre-trained LMs we consider (i) BERT (Devlin et al., 2019) and (ii) RoBERTa (Liu et al., 2019).

(b) Knowledge Encoding Layer: As depicted in Figure 3, the knowledge encoding layer is a Transformer-Block (Liu et al., 2018; Alt et al., 2019) as typically used in the decoder part of the transformer model of Vaswani et al. (2017). The core idea is that the model repeatedly encodes the given knowledge input over multiple layers (i.e., Transformer blocks), where each layer consists of masked multi-head self-attention followed by layer normalization and a feed-forward operation. Similar to the context input format, we concatenate the knowledge inputs as a sequence of tokens k_n = (k_{n1}, .. k_{nw}), where k_n is the knowledge used for input option x_n. In order to obtain the hidden knowledge representation we do the following:

$$h_{k_n^0} = k_n W_{ke} + W_{kp}, \quad h_{k_n^l} = tb(h_{k_n^{l-1}}), \forall l \in [1, L] \tag{1}$$

where W_{ke} is the token embedding matrix, W_{kp} the position embedding matrix, tb the transformer block, and L the number of transformer blocks.

(c) Reasoning Cell: The main intuition behind the reasoning cell is that given the context representation, the model should learn to emulate reasoning over the input using the knowledge representation obtained from the knowledge encoder. The Reasoning Cell is another transformer block, where the model repeatedly performs multi-head attention over the context and knowledge representations, and thus can iteratively refine the context representation. This capability is crucial for allowing the model to emulate complex reasoning steps through composition of various knowledge pieces. The multi-head attention function has three inputs: a query Q (context representation), key K and value

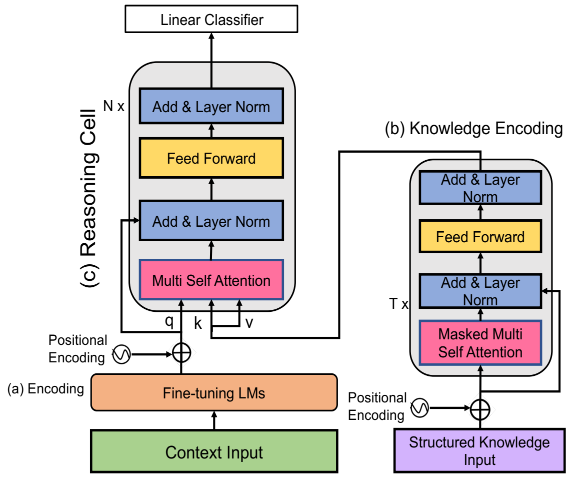

Figure 3: Overview of our Multi-Headed Knowledge Attention Model. It consist of three components (a) the *Context Encoding Layer* (b) the *Knowledge Encoding Layer*, and (c) the *Reasoning Cell*.

V (both knowledge representation). It relies on scaled dot-product attention

$$Q = h_{x_n} + W_{xp}, \quad a_{xk_n} = softmax(\frac{QK^T}{\sqrt{d_z}})V \tag{2}$$

where K = V = h_{k_n}, d_z the dimensionality of the input vectors representing the key and value, and W_{xp} is the position embedding. We project the output representations from the reasoning cell into logit (s) of size n (the number of output values) using a linear classifier. Finally, we compute the scores $y = max(s_i)$ where, $i = 1, .., n$. For CIP, where $n = 1$, we treat a logit score > 0 as predicting yes, otherwise the answer is no.

4.2 Applying the MHKA model to advanced Social Commonsense Reasoning Tasks

There are some key differences in how MHKA solves the two reasoning tasks:

(a) In the **abductive αNLI reasoning task**, the model must predict – given incomplete observations O_1 and O_2 – which of two hypotheses H_i is more plausible. For example: O_1: *Daniel wanted to buy a toy plane, but he didn't have any money*; O_2: *He bought his toy plane, and kept working so he could buy another*; correct H_i: *He opened a lemonade stand*. Here, the model needs to link O_2 back to O_1 using social inference knowledge relating to the H_i that best supports one of the sequences: O_1, H_i, O_2. In this case, the model

Task	Train	Dev	Test
αNLI	169654	1532	3059
CIP	12700	1008	1184

Table 3: Dataset Statistics: nb. of instances.

obtains the (encoded) input: O_1, H_i, O_2, and is tasked to predict the correct H_i, using available knowledge rules.[5]

(b) For **Counterfactual Invariance Prediction, CIP**, the model needs to decide whether for given a context C_{s_1,s_2,s_3}, under the assumption of a counterfactual s'_2, the given s_3 remains invariant or not. I.e., given: *Dotty was grumpy. Dotty called close friends to chat. She felt better afterwards.* and the counterfactual s'_2: *Dotty ate something bad* – can it still be true that *Dolly felt better afterwards*? Here our model gets as input the factual (s_2) and a counterfactual (s'_2) context: s_1, s_2, s_3 [SEP], s_1, s'_2, s_3 (cf. Table 2) and is tasked to predict whether or not s_3 remains true under the assumption s'_2. Again, the model needs to identify relevant knowledge to substantiate whether s_3 prevails given s_1 and s'_2.

Abduction *meets* Counterfactual Reasoning Clearly, when learning to judge whether s_3 holds true given both a factual (s_1, s_2) and counterfactual (s_1, s'_2) context, the CIP model learns how different events can or cannot lead to the very same factual event in a hypothetical reasoning task. Our intuition is that such a model effectively also acquires knowledge about what kinds of events can provide *evidence* for a given event, as is needed to perform *abduction*. Hence, we hypothesize that a model that has learned to understand and reason about counterfactual situations can also support abductive reasoning (i.e., finding the best explanation for an event). In our experiments, we test this hypothesis, and evaluate the performance of a model on the αNLI task, that we first train on CIP and then finetune it on the abductive inference task.

5 Experiments

Tasks and Settings. We apply our model to the two social reasoning tasks introduced in §2. We train models for each task using the input settings stated in Table 2. Data statistics is given in Table 3.

[5]Relevant knowledge from COMET2.0 here includes: [O_1: Daniel wanted to have money] $\rightarrow$ [H_i: Daniel wanted to make money, **Daniel then makes money**] $\rightarrow$ [O_2: Daniel needed to have money]. Clearly, H_i is supported by H_1: *He opened a lemonade stand.* So we can judge that the selected knowledge (partially) supports H_1.

We extract, for each event in each input sentence, social commonsense reasoning knowledge from COMET2.0, as detailed in §3. For the extraction process we use SRL as implemented in AllenNLP (Gardner et al., 2018).

Hyperparameter Details. In all models the Reasoning Cell and the Knowledge Encoder are both instantiated by a Transformer with 4 attention heads and depth=4. For each task, we select the hyperparameters that yield best performance on the dev set. Specifically, we perform a grid search over the hyperparameter settings with a learning rate in $\{1e\text{-}5, 2e\text{-}5, 5e\text{-}6\}$, a batch size in $\{4, 8\}$, and a number of epochs in $\{3, 5, 10\}$. Training is performed using cross-entropy loss. For evaluation, we measure accuracy. We report performance on the test sets by averaging results along with the variance obtained for 5 different seeds. See *Supplement* for details.

Baselines. We compare our model to the following baselines:
(a) *OpenAI-GPT* (Radford et al., 2018) is a multi-layer Transformer-Decoder based language model, trained with an objective to predict the next word.
(b) *Transformer Encoder* Model has the same architecture[6] as OpenAI-GPT without pre-training on large amounts of text.
(c) *BERT* (Devlin et al., 2019) is a LM trained with a masked-language modeling (MLM) and next sentence prediction objective, i.e., it is trained to predict words that are masked from the input.
(d) *RoBERTa* (Liu et al., 2019) has the same architecture as BERT, yet without next-sentence prediction objective. *RoBERTa-B(ase)* and *-L(arge)* were trained on more data and optimized carefully.
(e) *McQueen* (Mitra et al., 2019) proposed ways to infuse unstructured knowledge into pretrained language model (RoBERTa) to address the αNLI task. Mitra et al. (2019) used original *ROCStories* Corpus (Mostafazadeh et al., 2016) and Story Cloze Test that were used in creating αNLI dataset.
(f) $L2R^2$ (Learning to Rank for Reasoning) (Zhu et al., 2020) proposed to reformulate the αNLI task as a ranking problem. They use a learning-to-rank framework that contains a scoring function and a loss function.

6 Experimental Results

This section describes the experiments and results of our proposed model in different configurations.

[6]12-layer, 768-hidden, 12-heads

Model	Dev (%)	Test (%)
Majority ◇	50.8	–
GPT ◇	62.7	62.3
BERT -L ◇	69.1	68.9
McQueen (Mitra et al., 2019)	86.68	84.18
Concurrent Work		
$L2R^2$ (Zhu et al., 2020)	–	86.81
This work		
Transformer Enc. w/o LM-Pretraining	52.12	51.25
+ MHKA	54.96	53.91
RoBERTa-B	71.2±0.3	71.13±0.5
RoBERTa-B + MHKA	73.87±0.2	**74.17**±0.2
RoBERTa-L	85.06±0.7	84.48±0.7
RoBERTa-L + Joint Training	85.58±0.5	84.91±0.7
RoBERTa-L + MHKA	87.44±0.5	**87.12**±0.5
Human Perf.	–	91.4

Table 4: Results on αNLI dataset, ◇: as in Bhagavatula et al. (2020), L = Large, B = Base, excluding unpublished leaderboard submissions

Results on αNLI. Our experiment results for the α**NLI task** are summarized in Table 4. We compare performances of the following models: majority baseline, pre-trained LM baselines, and MHKA fine-tuned on RoBERTa-B(ase)/-L(arge). We observe consistent improvements of our MHKA method over RoBERTa-B (+3.04 percentage points, pp.) and RoBERTa-L (+2.64 pp.) on αNLI. Since MHKA uses RoBERTa to encode the input, this gain is mainly attributed to the use of knowledge and the multi-head knowledge attention technique. To better understand the impact of knowledge from pre-trained LMs, we trained a transformer encoder model *without* fine-tuning on a pretrained LM (see Table 4). Clearly, the overall performance of such a model drops considerably compared to the SOTA supervised models, but the improvement of MHKA by +2.84 points suggest that the impact of knowledge and reasoning obtained through multi-head knowledge attention is stable and independent from the power of LMs. Further, we compare our knowledge incorporation technique with *Joint Training*: this method uses pre-trained LMs to jointly encode both task-specific input and the knowledge ([CLS] (K)nowledge [SEP] (I)nput text). Table 4 shows that *Joint Training* yields limited improvement (+0.43 pp.) over the RoBERTa-L baseline – the intuitive reason being that the pretrained LMs were never trained on such structured knowledge.[7] However, our *MHKA model* shows a solid improvement of 2.64 pp. over the baseline. This suggests the impact of the *Multi-Head Knowledge Attention* integration technique.

[7]They also have a disadvantage when the length of context + knowledge increases, as this causes a bottleneck for computation on a GPU with limited memory (8-24GB).

Model	Input format	Dev%	Test%
RoBERTa-B	s_1, [SEP], s_2', [SEP], s_3	63.29	61.8
	s_1, s_2 [SEP] s_1, s_2'	57.44	58.9
	s_1, s_2, s_3 [SEP] s_1, s_2'	64.38	62.8
	s_1, s_2, s_3 [SEP] s_1, s_2' , s_3	66.66	67.98±0.5
+ MHKA	s_1, s_2, s_3 [SEP] s_1, s_2' , s_3	69.34	69.7±0.6
RoBERTa-L	s_1, s_2, s_3 [SEP] s_1, s_2' , s_3	72.4	71.95±0.6
+ MHKA	s_1, s_2, s_3 [SEP] s_1, s_2' , s_3	74.4	73.05±0.3
Human Perf.			**84.8**

Table 5: Results on Counterfactual Invariance Prediction (CIP).

Model	Dev	Test
RoBERTa-Large-αNLI	76.3	76.8
Transfer Learning	78.00	**79.04**
Transfer Learning + MHKA	78.6	**80.77**

Table 6: Impact of Counterfactual Invariance Prediction on αNLI. Training data size for αNLI is 8.5k (5%)

Low Resource Setting for αNLI. To better understand the impact of dataset scale on the performance of MHKA, and to test its robustness to data sparsity on αNLI, we investigate low-resource scenarios where we only use $\{1, 2, 5, 10, 100\}\%$ of the training data. Figure 4 shows constant advances of MHKA over both RoBERTa-Base and -Large. This result indicates the importance of knowledge in low-resource settings.

Results on CIP. Table 5 reports the results of our MHKA model on the CIP task, comparing to both RoBERTa baselines. As this is a new task, we also report the results of RoBERTa-Base with different input formats. We find that providing the model with the full sequence (s_1, s_2, s_3 [SEP] s_1, s_2' , s_3) gives best performance. By extending RoBERTa-Base and -Large with our MHKA reasoning component, we obtain an improvement of +1.7 and +1.1 percentage points, respectively.

CIP for Transfer Learning. We now test our hypothesis, discussed in §4.2, that a model trained on the CIP task can support the αNLI task. We first fine-tune two models: RoBERTa-L and the RoBERTa-L+MHKA model on the CIP task (using the hyperparameters for the CIP task, Table 5). As a transfer-learning method, we fine-tune these models on 5% of the training data for the αNLI task (using the hyperparameters for αNLI, Table 4) and report the results in Table 6 as "Transfer Learning"

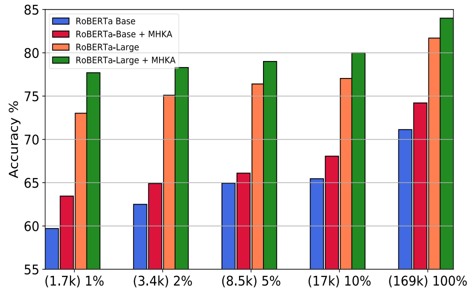

Figure 4: Accuracy for αNLI (Low Resource Setting)

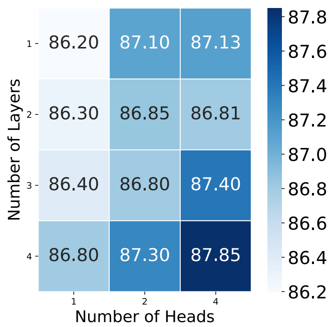

Figure 5: (a) Performance of MHKA model with different numbers of Heads and numbers of Layers.

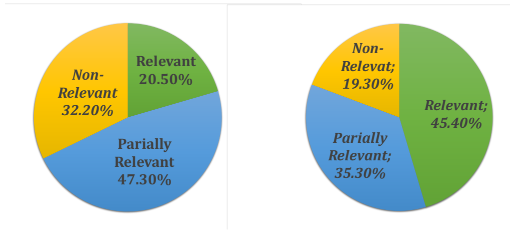

Figure 6: Human evaluation of the relevance of Knowledge Rules a) for 100 instances from the αNLI dev set and b) for the 56 (out of the 100) instances where the MHKA model predicted the correct hypothesis.

	all knowledge	**w/o irrelevant**	**w/o relevant + partially relevant**	**replacing relevant**
acc	56.2	57.6 (+1.4)	49.4 (−6.8)	45.05 (−11.2)
#	56	54 (−2)	20 (−36)	18 (−38)

Table 7: *row 1*: accuracy on 100 random instances from αNLI devset where the RoBERTa-L baseline fails; *row 2*: nb. of instances (#) correctly predicted by MHKA.

and "Transfer Learning + MHKA". Table 6 also reports the results for RoBERTa-L trained on 5% of the data of αNLI (called RoBERTa-L-αNLI).[8] We obtain a +2.84 pp. improvement over this baseline by applying the pre-trained CIP model on the αNLI task, and observe a further +1.73 pp. improvement (i.e., overall 3.97 points wrt. the baseline) with the stronger MHKA model. These results confirm our hypothesis, and show that learning to distinguish the outcomes of factual and counterfactual events can help the model to better perform abduction.

Ablation on Reasoning Cell. To give further insight into the factors for the model's capacity, we study the impact of the number of heads and layers in the reasoning cell. The left part of Figure 5(a) shows the performance of the MHKA model with different numbers of heads and layers. Note that the hidden dimensions of RoBERTa-Large is 1024 which is not divisible by 3, therefore we have 1, 2, and 4 as our attention heads. We observe that increasing the number of heads and layers improves the performance of the model. The intuitive explanation is that *multiple heads* help the model to focus on multiple knowledge rules and at the same time *multiple layers* help the model to recursively select the relevant knowledge rules.

[8]The training data size of αNLI is 14x larger than CIP. Therefore, in order to study the impact of CIP on αNLI, we made the training data size of CIP and αNLI comparable.

7 Analysis

Up to now, we have focused on performance analysis with different experimental settings and model ablations to analyze our model's capacities. Now, we turn to leveraging the fact that our model works with semi-structured knowledge in order to obtain deeper insight into its inner workings.

7.1 Quantitative Analysis.

Analysis on Knowledge Relevance. We conduct human evaluation to validate the effectiveness and relevance of the extracted social commonsense knowledge rules. We randomly select 100 instances from the αNLI dev set for which the RoBERTa-Large Baseline had failed, along with their gold labels and the extracted knowledge. Table 7 shows that MHKA correctly predicts 56 instances correctly. We asked two annotators to mark the knowledge rules that are relevant or partially relevant or irrelevant for each all 100 instances. The obtained answers yield that in 20.50% of cases the knowledge rules were relevant, in 47.30% of cases they were partially relevant (see Figure 6.a). Figure 6.b depicts the relevance of knowledge rules for instances that are *correctly predicted* by MHKA. The inter-annotator agreement had a Fleiss' κ=0.62.

Analysis of Model's Robustness. We then test the robustness of the models' performance by ma-

All knowledge	Removing relevant relation tuples	Removing relation tuples randomly
87.85	85.4 (−2.45)	86.9 (−0.95)

Table 8: Accuracy on αNLI (dev set)

nipulating the knowledge it receives for these instances in different ways: (a) we remove *irrelevant* and (b) *relevant* knowledge rules, (c) we manually change randomly selected rules from those that were found to be relevant by our annotators, and perturb them with artifacts. E.g., where annotators found that "PersonX's feelings" is relevant, we change the sentiment by choosing incorrect possible values from ATOMIC; for other relation types, we replace COMET's generated object with an antonym "PersonX wanted to be [nice → mean]", etc. We evaluate the effect of the perturbations i) on all 100 instances, and ii) on the 56 correctly predicted instances. Results are shown in Table 7. We see, for (a), a small improvement over the model results when using all knowledge, whereas for (b) and (c) an important performance drop occurs. For the 56 instances that MHKA resolves correctly, for (b) and (c) we find the same effect, but with a much more drastic drop in performance for (b) and (c).

This suggests that when the model is provided with relevant knowledge rules, it is able to utilize the knowledge well to perform the inference task.

In another test, we remove knowledge rules with relations which were found most relevant by our annotators (namely, *'PersonX's intent'*, *'PersonX's want'*, *'PersonX's need'*, *'effect on PersonX'*, *'effect on other'*, *'PersonX feels'*) (see *Supplement* for details). Table 8 reports the results on dev set.

We observe: (a) when we remove the *relevant* relational knowledge rules, the accuracy drops by 2.4 pp. suggesting that the model is benefitting from the knowledge rules. (b) when we remove knowledge rules *randomly*, the accuracy drop is minimal which shows the robustness of our model.

7.2 Qualitative Analysis.

Finally, we perform a study to better understand which knowledge rules were *"used or incorporated in the Reasoning Cell"* during the inference.

A case study. Figure 7 depicts an example from the *αNLI* task where we see the context at the top, and knowledge rules along with different scores below. The *Human scores* are annotated by the annotators where, 1.0 = Relevant, 0.50 = Partially relevant, 0.0 = Irrelevant. We also show the normalized attention scores over the structured knowledge rules[9]. We also measure a similarity score (using dot product) between the final representation of the Reasoning cell and different knowledge rules. Intuitively, we expect that relevant knowledge rules should be incorporated in the final representation of the Reasoning cell, and therefore, should have a higher similarity score compared to irrelevant knowledge rules. Figure 7, illustrates one such example where we see that some relevant knowledge (judged by annotators) – *"He gained 20 pounds in one month ⟨xIntent⟩ He wanted to lose weight"*, and *"He would stop eating fast food ⟨xWant⟩ he wants to lose weight"* – are highly attended, and scored higher in similarity measure compared to others, indicating that the Reasoning Cell incorporated these knowledge rules. To study this further, we randomly selected 10 instances from the *αNLI* dev set along with the knowledge rules. We found for 7 out of 10 instances that the MHKA model gave higher similarity scores to relevant or partially relevant knowledge rules than to irrelevant ones.

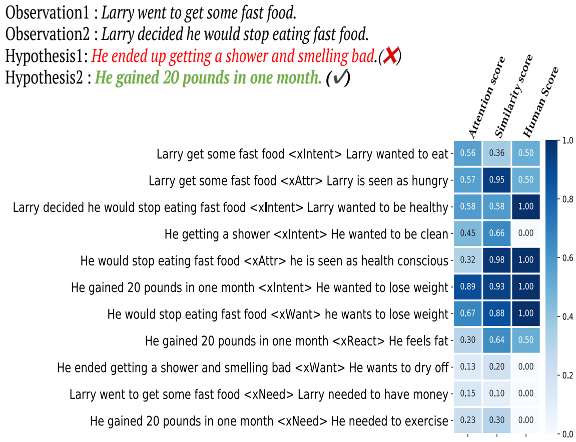

Figure 7: Comparing relevance scores of knowledge.

8 Related Work

Social Commonsense Knowledge Teaching machines to reason about daily events with commonsense knowledge has been an important component for natural language understanding (McCarthy, 1959; Davis and Marcus, 2015; Storks et al., 2019). Given the growth of interest among researchers in commonsense reasoning, a large body of work has been focused on learning commonsense knowl-

[9]Note that we do not consider the attention maps as explanations. We assume that attention exhibits an intuitive interpretation of the model's inner workings.

edge representations (Lenat, 1995; Espinosa and Lieberman, 2005; Speer et al., 2017; Tandon et al., 2017). In this work, we address social commonsense reasoning, where knowledge about events and its implications is crucial. Rashkin et al. (2018) (Event2Mind) proposed a knowledge resource for commonsense inference about people's intentions and reactions in everyday events. Later, Sap et al. (2019) (ATOMIC) extended the Event2Mind resource with substantially more events, and with nine dimensions (*If-then* relation types) per event. There has also been work on automatically acquiring commonsense knowledge (Li et al., 2016; Bosselut et al., 2019; Malaviya et al., 2020). Recently, Nasrin Mostafazadeh (2020) introduced a large-scale dataset (GLUCOSE) capturing ten dimensions of causal explanation (implicit commonsense knowledge) in a narrative context. However, learning to reason over such event-based semi-structured knowledge is still a challenging task. In this work, we propose a model which learns to imitate reasoning using such structured knowledge.

Commonsense Reasoning (CR): There is a large body of research on commonsense reasoning over natural language text (Levesque et al., 2012; Bowman et al., 2015; Zellers et al., 2019; Trichelair et al., 2019; Becker et al., 2020). We discuss the ones most related to our work. Earlier works sought to utilize rule-based reasoning or hand-crafted features (Sun, 1995; Gupta and Hennacy, 2005). With the increase in size of commonsense knowledge bases (Suchanek et al., 2007; Speer et al., 2017) researchers started utilizing them to help models perform commonsense reasoning (Schüller, 2014; Liu et al., 2017). Recently, there have been attempts to leverage pre-trained language models to learn and perform commonsense inference, and they achieved state-of-the-art results (Radford et al., 2018; Trinh and Le, 2018; Kocijan et al., 2019; Radford et al., 2019). Our model takes advantage of both pre-trained LMs and structured knowledge, which allows us to inspect the reasoning process. We also demonstrate that our model shows strong performance for different, and finely structured tasks in abductive and counterfactual reasoning.

Structured Commonsense Knowledge in Neural Systems: Different approaches have been proposed to extract and integrate external knowledge into neural models for various NLU tasks such as reading comprehension (RC) (Xu et al., 2017; Mihaylov and Frank, 2018; Weissenborn et al., 2018), question answering (QA) (Xu et al., 2016; Tandon et al., 2018; Wang et al., 2019), etc. Recently, many works proposed different ways to extract knowledge from static knowledge graphs (KGs). Most notable are ones that extract subgraphs from KGs using either heuristic methods (Bauer et al., 2018) or graph-based ranking methods (Paul and Frank, 2019; Paul et al., 2020), or else utilize knowledge graph embeddings (Lin et al., 2019) to rank and select relevant knowledge triples or paths.

Similar to Bosselut and Choi (2019) and Shwartz et al. (2020), in this work we generate contextually relevant knowledge using language models trained on KGs. With the increase in performance of transformer-based models there has been a shift from RNN-based neural models to pre-trained LMs. Incorporating extracted knowledge using attention mechanism (single dot product) has become a standard procedure. However, we propose a multi-head attention model that can recursively select multiple generated structured knowledge rules, and also allows inspection by analyzing the used knowledge.

9 Conclusion

In this work, we propose a new *multi-head knowledge attention model* to incorporate semi-structured social commonsense knowledge. We show that our model improves over state-of-the-art LMs on two complex commonsense inference tasks. Besides the improvement i) we demonstrate a correlation between abduction and counterfactual reasoning in a narrative context, based on the newly proposed task of counterfactual invariance prediction, which we apply to support abductive inference. Importantly, ii) we confirm the reasoning capacity of our model by perturbing and adding noise to the knowledge, and performing model inspection using manually validated knowledge rules. In future work, we aim to deeper investigate compositional effects of inferencing, such as the interaction of socially grounded and general inferential knowledge.

Acknowledgements

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1. We thank our annotators for their valuable annotations. We also thank NVIDIA Corporation for donating GPUs used in this research.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230.

Maria Becker, Ioana Hulpus, Debjit Paul, Juri Opitz, Jonathan Kobbe, Heiner Stuckenschmidt, and Anette Frank. 2020. Explaining Arguments with Background Knowledge – Towards Knowledge-based Argumentation Analysis . *Datenbank Spektrum (Special Issue: Argumentation Intelligence)*, 20:131–141.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

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

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58:92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kai Epstude and Neal J Roese. 2008. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192.

Jose Espinosa and Henry Lieberman. 2005. Eventnet: Inferring temporal relations between commonsense events. In *Mexican International Conference on Artificial Intelligence*, pages 61–69. Springer.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Rakesh Gupta and Ken Hennacy. 2005. Commonsense reasoning about task instructions. In *AAAI-05 Workshop on modular construction of human-like intelligence. Pittsburgh, PA, July*, volume 10, pages 05–08.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

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

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2822–2832.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017. Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *IJCAI*, pages 2344–2350.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

John McCarthy. 1959. Programs with common sense. *in Proceedings of the Teddington Conference on the Mechanization of Thought Processes, London: Her Majesty's Stationary Office*, pages 75–91.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *arXiv preprint arXiv:1909.08855*.

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

Lori Moon David Buchanan Lauren Berkowitz Or Biran Jennifer Chu-Carroll Nasrin Mostafazadeh, Aditya Kalyanpur. 2020. Glucose: Generalized and contextualized story explanations. In *EMNLP*.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3671–3681, Minneapolis, Minnesota. Association for Computational Linguistics.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative Relation Classification with Background Knowledge. In *Proceedings of the 8th International Conference on Computational Models of Argument (COMMA 2020)*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 319–330. Computational Models of Argument.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *2019 Conference on Empirical Methods in Natural Language Processing.*, Hongkong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Neal J Roese and Mike Morrison. 2009. The psychology of counterfactual thinking. *Historical Social Research/Historische Sozialforschung*, pages 16–26.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 3027–3035.

Peter Schüller. 2014. Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 358–367. AAAI Press.

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv preprint arXiv:1904.05255*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, , and Yejin Choi. 2020. Unsupervised Commonsense Question Answering with Self-talk. In *EMNLP*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shane Storks, Qiaozi Gao, and Joyce Yue Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *ArXiv*, abs/1904.01172.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Ron Sun. 1995. Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, 75(2):241–295.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*.

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 57–66, Brussels, Belgium. Association for Computational Linguistics.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*, pages 115–120.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3382–3387, Hong Kong, China. Association for Computational Linguistics.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215.

Dirk Weissenborn, Tomáš Kočiskỳ, and Chris Dyer. 2018. Dynamic integration of background knowledge in neural NLU systems.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on Freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336, Berlin, Germany. Association for Computational Linguistics.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 3506–3513. IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Yunchang Zhu, Liang Pang, Yanyan Lan, and Xueqi Cheng. 2020. L2r: Leveraging ranking for abductive reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1961–1964, New York, NY, USA. Association for Computing Machinery.