Commonsense Inference in Natural Language Processing (COIN) — Shared Task Report

Simon Ostermann
Saarland University / Nuance Communications
simono@coli.uni-saarland.de

Sheng Zhang
Johns Hopkins University
zsheng2@jhu.edu

Michael Roth
Stuttgart University/ Saarland University
rothml@ims.uni-stuttgart.de

Peter Clark
Allen Institute for Artificial Intelligence
peterc@allenai.org

Abstract

This paper reports on the results of the shared tasks of the COIN workshop at EMNLP–IJCNLP 2019. The tasks consisted of two machine comprehension evaluations, each of which tested a system's ability to answer questions/queries about a text. Both evaluations were designed such that systems need to exploit commonsense knowledge, for example, in the form of inferences over information that is available in the common ground but not necessarily mentioned in the text. A total of five participating teams submitted systems for the shared tasks, with the best submitted system achieving 90.6% accuracy and 83.7% F1-score on task 1 and task 2, respectively.

1 Introduction

Due to the rise of powerful pre-trained word and sentence representations, automated text processing has come a long way in recent years, with systems that perform even better than humans on some datasets (Rajpurkar et al., 2016a). However, natural language understanding also involves complex challenges. One important difference between human and machine text understanding lies in the fact that humans can access commonsense knowledge while processing text, which helps them to draw inferences about facts that are not mentioned in a text, but that are assumed to be common ground.

(1) *Max:* "It's 1 pm already, I think we should get lunch."
Dustin: "Let me get my wallet."

Consider the conversation in Example 1: Max will not be surprised that Dustin needs to get his wallet, since she knows that *paying* is a part of *getting lunch*. Also, she knows that a wallet is needed for paying, so Dustin needs to get a wallet for lunch. This is part of the commonsense knowledge about getting lunch and should be known by both persons. For a computer system, inferring such unmentioned facts is a non-trivial challenge. The workshop on Commonsense Inference in NLP (COIN) is focused on such phenomena, looking at models, data, and evaluation methods for commonsense inference.

This report summarizes the results of the COIN shared tasks, an unofficial extension of the SemEval 2018 shared task 11, *Machine Comprehension using Commonsense Knowledge* (Ostermann et al., 2018b). The tasks aim to evaluate the commonsense inference capabilities of text understanding systems in two settings: Commonsense inference in everyday narrations (task 1) and commonsense inference in news texts (task 2). Framed as machine comprehension evaluations, the datasets used for both tasks contain challenging reading comprehension questions asking for facts that are not explicitly mentioned in the given reading texts.

Several teams participated in the shared tasks and submitted system description papers. All systems are based on Transformer architectures (Vaswani et al., 2017), some of them explicitly incorporating commonsense knowledge resources, whereas others only use pretraining on other machine comprehension data sets. The best submitted system achieves 90.6% accuracy and 83.7% F1-score on task 1 and task 2, respectively. Still, there are cases that remain elusive: Humans outperform this system by a margin of 7% (task 1) and 8% (task 2). Our results indicate that while Transformer models are able to perform extremely well on the data used in our shared task, there are still some remaining cases demonstrating that human level is not achieved yet. Still, we believe that our results also imply the need for more challenging data sets. In particular, we need data sets that

Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 66–74
Hongkong, China, November 3, 2019.

make it harder to benefit from redundancy in the training data or large-scale pretraining on similar domains.

In the following, we briefly describe the data sets (§2), baselines and evaluation metrics of the shared tasks (§3) and we present a summary of the participating systems (§4), their results (§5) as well as a discussion thereof (§6).

2 Data and Tasks

Text understanding systems are often evaluated by means of a reading comprehension task, which is also referred to as machine (reading) comprehension (MC). The central idea is that a system has to process a text and then find a correct answer to a question that is asked on the text. Our shared tasks follow this paradigm and use machine comprehension settings to evaluate a model's capability to perform commonsense inferences. In contrast to most existing MC datasets, the two datasets that are used for our shared tasks, *MCScript2.0* (Ostermann et al., 2019) and ReCoRD (Zhang et al., 2018), are focused on questions that cannot be answered from the text alone, but that require a model to draw inference over unmentioned facts.

(2) *Text:* Camping is one of my favorite summer vacations. (...) Once I have all my gear and clothing I'll pack it into my car, making sure to leave room for myself, my dog and anything my friends want to bring. And then we are ready for our camping vacation.
Question: What do they put the drinks in?
a. Cooler
b. Sleeping bag

Example 2 illustrates the main idea of the shared tasks. It shows a reading text from MCScript2.0, together with a question and two candidate answers. For a human, it is trivial to find that the drinks are put into a cooler rather than the sleeping bag. This information is however not mentioned in the text, so a machine needs to have the capability to infer this fact from commonsense knowledge.

The reading texts of MCScript2.0 are narrations about everyday activities (task 1). Due to its domain, MCScript2.0 has a focus on evaluating script knowledge, i.e. knowledge about the events and participants of such everyday activities (Schank and Abelson, 1975). Task 2 utilizes the ReCoRD corpus (Zhang et al., 2018), which contains news texts, a more open domain. The inferences that are required for finding answers to the questions in ReCoRD are thus of a more general type.

2.1 Task 1: Commonsense Inference in Everyday Narrations

MCScript2.0 is a reading comprehension data set comprising 19,821 questions on 3,487 texts. Each of the questions has two answer candidates, one of which is correct. Questions in the data were annotated for reasoning types, i.e. according to whether the answer to a question can be found in the text or needs to be inferred from commonsense knowledge. Roughly half of the questions do require inferences over commonsense knowledge.

The texts in MCScript2.0 are short narrations (164.4 tokens on average) on a total of 200 different everyday activities. All texts were crowdsourced on Amazon Mechanical Turk[1], by asking crowd workers to tell a story about one of the 200 scenarios *as if talking to a child* (Modi et al., 2016; Ostermann et al., 2018a), resulting in simple texts which explicitly mention many details of a scenario. In the question collection, which was also conducted via crowdsourcing, turkers were then asked to write questions about noun or verb phrases that were highlighted in the texts. After collecting questions, the sentences containing the noun or verb phrases were deleted from the texts. During the answer collection, crowd workers thus had to infer the information required for finding an answer from background knowledge. Five turkers wrote correct and incorrect answer candidates for each question, and the most difficult incorrect candidates were selected via adversarial filtering (Zellers et al., 2018).

For our shared task, we use the same data split as Ostermann et al. (2019): 14,191 questions on 2,500 texts for the training set, 2,020 questions on 355 texts for the development set and 3,610 questions on 632 texts for the test set. All texts for five scenarios were reserved for the test set only to increase difficulty.

2.2 Task 2: Commonsense Inference in News Articles

ReCoRD is a large-scale dataset for reading comprehension, which consists of over 120,000 ex-

[1] https://www.mturk.com/

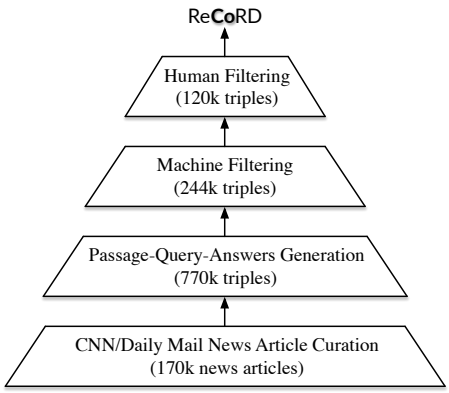

Figure 1: ReCoRD data collection procedure.

	Train	Dev.	Test	Overall
queries	100,730	10,000	10,000	120,730
unique passages	65,709	7,133	7,279	80,121
passage vocab.	352,491	93,171	94,386	395,356
query vocab.	119,069	30,844	31,028	134,397
tokens / passage	169.5	168.6	168.1	169.3
entities / passage	17.8	17.5	17.3	17.8
tokens / query	21.3	22.1	22.2	21.4

Table 1: Statistics of ReCoRD

amples, most of which require commonsense reasoning. ReCoRD was collected in a four-stage process (Figure 1): (1) curating CNN/Daily Mail news articles, (2) generating passage-query-answers triples based on the news articles, (3) filtering out the queries that can be easily answered by state-of-the-art machine comprehension (MC) models, and (4) filtering out the queries ambiguous to human readers. All named entities in the passages are possible answers to the queries. Table 1 summarizes the data statistics.

3 Shared Task Setup

The baselines for our shared tasks were adapted from Ostermann et al. (2019) and Zhang et al. (2018), respectively.

3.1 Task 1 Baselines

Following Ostermann et al. (2019), we present results of three baseline models.

Logistic Regression Model. Merkhofer et al. (2018) presented a logistic regression classifier for the SemEval 2018 shared task 11, which used simple overlap features and word patterns on *MCScript*, a predecessor of the dataset used for this task. Their model outperformed many neural networks in spite of its simplicity.

Attentive Reader. The second baseline model is an attentive reader network (Hermann et al., 2015). GRU units (Cho et al., 2014) are used to process text, question and answer. A question-aware text representation is computed based on a bilinear attention function, which is then combined with a GRU-based answer representation for prediction. For details, we refer to Ostermann et al. (2019), Ostermann et al. (2018a) and Chen et al. (2016)

TriAN. As last model, we use the three-way attentive network (TriAN) (Wang et al., 2018), a recurrent neural network that scored the first place in the SemEval 2018 task. They use LSTM units (Hochreiter and Schmidhuber, 1997), several attention functions, and self attention to compute representations for text, question and answer. ConceptNet (Speer et al., 2017), a large commonsense knowledge base containing thousands of entities and commonsense relations between them, is used to enhance text representations with commonsense information, by computing relation embeddings and appending them to the text representations. For more information we refer to Wang et al. (2018).

3.2 Task 2 Baselines

We present five baselines for ReCoRD:

BERT (Devlin et al., 2019) is a new language representation model. Recently fine-tuning the pre-trained BERT with an additional output layer has created state-of-the-art models on a wide range of NLP tasks. We formalized ReCoRD as an extractive QA task like SQuAD, and then reused the fine-tuning script for SQuAD to fine-tune BERT for ReCoRD.

KT-NET (Yang et al., 2019a) employs an attention mechanism to adaptively select desired knowledge from knowledge bases, and then fuses selected knowledge with BERT to enable context- and knowledge-aware predictions for machine reading comprehension.

SAN (Liu et al., 2018) is a top-ranked MC model. It shares many components with other MC models, and employs a stochastic answer module. As we used SAN to filter out queries in the data collection, it is necessary to verify that the collected queries are hard for not only SAN but also other MC architectures.

Rank	Team Name	Architecture	Commonsense	Other Resources	Tasks
1	PSH–SJTU	Transformer (XLNet)	-	RACE, SWAG	1, 2
2	IIT-KGP	Transformer (BERT + XLNet)	-	RACE	1
3	BLCU-NLP	Transformer (BERT)	-	ReCoRD, RACE	1
4	JDA	Transformer (BERT)	ConceptNet, Atomic, Webchild	Wikipedia	1
5	KARNA	Transformer (BERT)	ConceptNet	-	1

Table 2: Overview of participating systems

DocQA (Clark and Gardner, 2018) is a strong baseline model for extractive QA. It consists of components such as bi-directional attention flow (Seo et al., 2016) and self-attention, both of which are widely used in MC models. We also evaluated a variant of DocQA with ELMo (Peters et al., 2018) to analyze the impact of ELMo on this task.

Random Guess acts as the lower bound of the evaluated models, which randomly picks a named entity from the passage as the answer. The reported results are averaged over 5 runs.

3.3 Evaluation

Task 1. The evaluation measure for task 1 is accuracy, computed as the number of correctly answered questions divided by the number of all questions. We also report accuracy values on questions that crowd workers explicitly annotated as requiring commonsense as well as performance on the five held-out scenarios.

Task 2. We use two evaluation metrics, EM and F1, similar to those used by SQuAD (Rajpurkar et al., 2016b). Exact Match (EM) measures the percentage of predictions that match a reference answer exactly. (Macro-averaged) F_1 measures the average overlap between model predictions and reference answers. For computing F_1, we treat prediction and reference answers as bags of tokens. We take the maximum F_1 over all reference answers for a given query, and then average over all queries.

4 Participants

In total, five teams submitted systems in task 1, and one team participated in task 2. All submitted models were neural networks, and all made use of pretrained Transformer language models such as *BERT* (Devlin et al., 2019). The participants used a wide range of external corpora and resources to augment their models, ranging from other machine comprehension data sets such as *RACE* (Lai et al., 2017) or *MCScript* (Ostermann et al., 2018a), up to commonsense knowledge databases such as *ConceptNet* (Speer et al., 2017), *WebChild* (Tandon et al., 2017) or ATOMIC (Sap et al., 2019). Table 2 gives a summary of the participating systems.

- **PSH–SJTU** (Li et al., 2019) participated in both tasks with a Transformer model based on *XLNet* (Yang et al., 2019b). For task 1, they pretrain the model in several steps, first on the RACE data (Lai et al., 2017) and then on SWAG (Zellers et al., 2018). For task 2, they do not conduct specific pretraining steps, but implement a range of simple rule-based answer verification strategies to verify the output of the model.

- **IIT-KGP** (Sharma and Roychowdhury, 2019) present an ensemble of different pretrained language models, namely BERT and XLNet. Both models are pretrained on the RACE data (Lai et al., 2017), and their output is averaged for a final prediction.

- **BLCU-NLP** (Liu et al., 2019) use a Transformer model based on BERT, which is fine-tuned in two stages: they first tune the BERT-based language model on the RACE and ReCoRD datasets and then (further) train the model for the actual machine comprehension task.

- **JDA** (Da, 2019) use three different knowledge bases, namely ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019) and WebChild (Tandon et al., 2017). They extract

relevant edges from the knowledge bases and compute relation embeddings, which are combined with BERT-based word representations with a diadic multiplication operation.

- **KARNA** (Jain and Singh, 2019) use a BERT model, but they enhance the text representation with edges that are extracted from ConceptNet. Following Wang et al. (2018), they extract relations between words in the text and the question/answer, and append them to the text representation. Instead of computing relational embeddings, they append a specific string that describes the relation.

5 Results

Table 3 shows the performance of the participating systems and the baselines on the task 1 data. We tested for significance using a pairwise approximate randomization test (Yeh, 2000) over questions. Except for the two top scoring systems, each system performs significantly better than the next in rank. All systems significantly outperform the baselines. All systems show a lower performance on commonsense-based questions as compared to the average on all questions, with the difference for the two top-scoring systems being smallest. Surprisingly, all models are able to perform better on the questions from held-out scenarios as compared to their performance on all questions. This indicates that all models are able to generalize well from the training material.

Table 5 shows the systems' performance on single question types for task 1. Question types are determined automatically, as described in (Ostermann et al., 2019). As can be seen, both top-scoring systems perform well over all different question types, indicating that both systems are able to model a wide range of phenomena. Interestingly, *when* questions seem to be the most challenging question type for all systems, indicating difficulties when it comes to model event ordering information. Also, *where* questions seem to be challenging, at least for some systems.

Table 4 shows EM (%) and F_1 (%) of human performance, the PSH-SJTU system as well as baselines on the development and test sets of task 2. Compared with the best baseline, KT-NET (Yang et al., 2019a), PSH-SJTU achieves significantly better scores. On the hidden test set, they improve EM by 10.08%, and F_1 by 8.98%.

#	Team Name	acc	acc_{cs}	acc_{OOD}
1	PSH–SJTU	**0.906**	**0.903**	0.915
2	IIT-KGP	0.905*	0.894	**0.931**
3	BLCU-NLP	0.842*	0.812	0.838
4	JDA	0.807*	0.775	0.796
5	KARNA	0.733*	0.697	0.729
-	TriAN	0.715	0.666	0.673
-	Attentive Reader	0.651	0.634	0.619
-	Logistic	0.608	0.562	0.544
-	*Human*	0.97		

Table 3: Performance of participating systems and baselines for **task 1**, in total (acc), on commonsense-based questions (acc_{cs}), and on out-of-domain questions that belong to the five held-out scenarios (acc_{OOD}). Significant differences in results between two adjacent lines are marked by an asterisk (* $p<0.05$) in the upper line. The best model performance per column is marked in **bold print**.

	Dev.		Test	
	EM(%)	F_1(%)	EM(%)	F_1(%)
Human	**91.28**	**91.64**	**91.31**	**91.69**
PSH-SJTU	**82.72**	**83.38**	**83.09**	**83.74**
KT-NET	71.60	73.61	73.01	74.76
BERT-Large	66.11	68.49	67.61	70.01
SAN	48.86	50.08	50.43	51.41
DocQA	44.13	45.39	45.44	46.65
Random	18.41	19.06	18.55	19.12

Table 4: Performance (EM and F_1) of human, participating systems and baselines for **task 2**.

Consequently, PSH-SJTU has reduced the gap between human and machine performance, with human performance being only 8% higher than PSH-SJTU.

6 Discussion

Pretrained Transformer language models. A main finding of our shared tasks is that large pretrained Transformer language models such as BERT or XLNet perform well even on challenging commonsense inference data. Strikingly, all models generalize well, as can be seen from the good performance on held-out scenarios. On task 1, XLNet-based systems perform best. The difference to the models purely based on BERT

#	Team Name	what	when	where	who	how
1	PSH–SJTU	0.918	0.891	0.890	0.921	0.890
2	IIT-KGP	0.915	0.897	0.890	0.921	0.925
3	BLCU-NLP	0.874	0.800	0.815	0.857	0.870
4	JDA	0.844	0.777	0.744	0.794	0.829
5	KARNA	0.755	0.683	0.734	0.750	0.788
-	TriAN	0.749	0.647	0.712	0.730	0.801
-	AttentiveReader	0.700	0.578	0.620	0.659	0.726
-	Logistic	0.644	0.546	0.573	0.663	0.685

Table 5: Performance of participating systems and baselines for **task 1** on the 5 most common question types.

can mostly be attributed to the performance on commonsense-based questions: While the performance of XLNet-based models on such questions is almost on par with their average performance, models based on BERT underperform on commonsense questions. An interesting observation was made by Li et al. (2019), who found that including WordNet into a BERT model boosts performance, while there is no such boost for an XLNet model. This seems to indicate that XLNet is able to cover (at least partially) some form of lexical background knowledge, as encoded in WordNet, without explicitly requiring access to such a resource.

Still, when inspecting questions that were not answered correctly by the best scoring model, we found a large number of commonsense-based *when* questions that ask for the typical order of events. This indicates that XLNet-based models are only to a certain extent able to model complex phenomena such as temporal order.

Commonsense knowledge databases. Only two participants made use of commonsense knowledge, in the form of knowledge graphs such as ConceptNet. Both participants conducted ablation tests indicating the importance of including commonsense knowledge. In comparison to ATOMIC and WebChild, Da (2019) report that ConceptNet is most beneficial for performance on the task 1 data, which can be explained with its domain: The OMCS (Singh et al., 2002) data are part of the ConceptNet database, and OMCS scenarios were also used to collect texts for the task 1 data.

All in all, powerful pretrained models such as XLNet still outperform approaches that make use of structured knowledge bases, which indicates that they are (at least to some extent) capable of performing commonsense inference without explicit representations of commonsense knowledge.

Pretraining and finetuning on other data. Several participants reported effects of pretraining/finetuning their models on related tasks. For instance, Liu et al. (2019) experimented with different pretraining corpora and found results to be best when pretraining the encoder of their BERT model on RACE and ReCoRD. Similarly, Li et al. (2019) report improved results when using larger data sets from other reading comprehension (RACE) and commonsense inference tasks (SWAG) for training before fine-tuning the model with the actual training data from the shared task.

7 Related Work

Evaluating commonsense inference via machine comprehension has recently moved into the focus of interest. Existing datasets cover various domains:

Web texts. *TriviaQA* (Joshi et al., 2017) is a corpus of webcrawled trivia and quiz-league websites together with evidence documents from the web. A large part of questions requires a system to make use of factual commonsense knowledge for finding an answer. *CommonsenseQA* (Talmor et al., 2018) consists of 9,000 crowdsourced multiple-choice questions with a focus on relations between entities that appear in ConceptNet (Speer et al., 2017). Evidence documents were webcrawled based on the question and added after the crowdsourcing step.

Fictive texts. *NarrativeQA* (Kočiský et al., 2018) provides full novels and other long texts as evidence documents and contains approx. 30 crowdsourced questions per text. The questions

require a system to understand the whole plot of the text and to conduct many successive complicated inference steps, under the use of various types of background knowledge.

News texts. NewsQA (Trischler et al., 2017) provides news texts with crowdsourced questions and answers, which are spans of the evidence documents. The question collection procedure for NewsQA resulted in a large number of questions that require factual commonsense knowledge for finding an answer.

Other tasks. There have been other attempts at evaluating commonsense inference apart from machine comprehension. One example is the Story cloze test and the *ROC* dataset (Mostafazadeh et al., 2016), where systems have to find the correct ending to a 5-sentence story, using different types of commonsense knowledge. *SWAG* (Zellers et al., 2018) is a natural language inference dataset with a focus on difficult commonsense inferences.

8 Conclusion

This report presented the results of the shared tasks at the Workshop for Commonsense Inference in NLP (COIN). The tasks aimed at evaluating the capability of systems to make use of commonsense knowledge for challenging inference questions in a machine comprehension setting, on everyday narrations (task 1) and news texts (task 2). In total, 5 systems participated in task 1, and one system participated in task 2. All submitted models were Transformer models, pretrained with a language modeling objective on large amounts of textual data. The best system achieved 90.6% accuracy and 83.7% F1-score on task 1 and 2, respectively, leaving a gap of 7% and 8% to human performance.

The results of our shared tasks suggest that existing models cover a large part of the commonsense knowledge required for our data sets in the domains of narrations and news texts. This does however not mean that commonsense inference is solved: We found a range of examples in our data that are not successfully covered. Furthermore, data sets such as *HellaSWAG* (Zellers et al., 2019) show that commonsense inference tasks can be specifically tailored to be hard for Transformer models. We believe that modeling true language understanding requires a shift towards text types and tasks that test commonsense knowledge going beyond information that can be obtained by exploiting the redundancy of large-scale corpora and/or pretraining on related tasks.

References

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2358–2367.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855. Association for Computational Linguistics.

Jeffrey Da. 2019. Jeff Da at COIN - Shared Task. In *Proceedings of the 2019 EMNLP Workshop COIN: Commonsense Inference in NLP*.

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.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yash Jain and Chinmay Singh. 2019. KARNA at COIN - Shared Task: Bidirectional Encoder Representations from Transformers with relational knowledge for machine comprehension with common sense . In *Proceedings of the 2019 EMNLP Workshop COIN: Commonsense Inference in NLP*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Xiepeng Li, Zhexi Zhang, Wei Zhu, Yuan Ni, Peng Gao, Junchi Yan, and Guotong Xie. 2019. Pingan Smart Health and SJTU at COIN - Shared Task: Utilizing Pre-trained Language Models and Commonsense Knowledge in Machine Reading Tasks. In *Proceedings of the 2019 EMNLP Workshop COIN: Commonsense Inference in NLP*.

Chunhua Liu, Shike Wang, Bohan Li, and Dong Yu. 2019. BLCU-NLP at COIN - Shared Task: Stagewise Fine-tuning BERT for Commonsense Inference in Everyday Narrations. In *Proceedings of the 2019 EMNLP Workshop COIN: Commonsense Inference in NLP*.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704. Association for Computational Linguistics.

Elizabeth M. Merkhofer, John Henderson, David Bloom, Laura Strickhart, and Guido Zarrella. 2018. MITRE at SemEval-2018 Task 11: Commonsense Reasoning without Commonsense Knowledge. In *Proceedings of the 12th International Workshop on Semantic Evaluations (SemEval-2018)*, pages 1078–1082.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3485–3493. European Language Resources Association (ELRA).

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.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018a. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3567–3574.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018b. SemEval-2018 Task 11: Machine Comprehension using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A Machine Comprehension Corpus Focused on Script Events and Participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 103–117.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016a. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016b. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Roger C Schank and Robert P Abelson. 1975. Scripts, Plans, and Knowledge. In *Proceedings of the 4th international joint conference on Artificial intelligence-Volume 1*, pages 151–157. Morgan Kaufmann Publishers Inc.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Prakhar Sharma and Sumegh Roychowdhury. 2019. IIT-KGP at COIN - Shared Task: Using pre-trained Language Models for modeling Machine Comprehension. In *Proceedings of the 2019 EMNLP Workshop COIN: Commonsense Inference in NLP*.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. In *On the move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 4444–4451.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. *arXiv preprint arXiv:1811.00937*.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *The 55th Annual Meeting of the Association for Computational Linguistics*, pages 115–120. ACL.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

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.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018. Yuanfudao at SemEval-2018 Task 11: Three-way Attention and Relational Knowledge for Commonsense Machine Comprehension. In *Proceedings of the 12th International Workshop on Semantic Evaluations (SemEval-2018)*, pages 758–762.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019a. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Alexander Yeh. 2000. More accurate tests for the statistical significance of the result differences. In *Proc. 17th International Conf. on Computational Linguistics*, pages 947–953, Saarbruken, Germany.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

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*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv preprint arXiv:1810.12885*.