Introductory Tutorial:
Commonsense Reasoning for Natural Language Processing

Maarten Sap [1] **Vered Shwartz** [1,2] **Antoine Bosselut** [1,2] **Yejin Choi** [1,2] **Dan Roth** [3]
[1] Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA
[2] Allen Institute for Artificial Intelligence, Seattle, WA, USA
[3] Department of Computer and Information Science, University of Pennsylvania
{msap, vereds, antoineb, yejin} @cs.washington.edu, danroth@seas.upenn.edu

1 Introduction

Commonsense knowledge, such as knowing that "bumping into people annoys them" or "rain makes the road slippery", helps humans navigate everyday situations seamlessly (Apperly, 2010). Yet, endowing machines with such human-like commonsense reasoning capabilities has remained an elusive goal of artificial intelligence research for decades (Gunning, 2018).

Commonsense knowledge and reasoning have received renewed attention from the natural language processing (NLP) community in recent years, yielding multiple exploratory research directions into automated commonsense understanding. Recent efforts to acquire and represent common knowledge resulted in large knowledge graphs, acquired through extractive methods (Speer et al., 2017) or crowdsourcing (Sap et al., 2019a). Simultaneously, a large body of work in integrating reasoning capabilities into downstream tasks has emerged, allowing the development of smarter dialogue (Zhou et al., 2018) and question answering agents (Xiong et al., 2019).

Recent advances in large pretrained language models (e.g., Devlin et al., 2019; Liu et al., 2019b), however, have pushed machines closer to human-like understanding capabilities, calling into question whether machines should directly model commonsense through symbolic integrations. But despite these impressive performance improvements in a variety of NLP tasks, it remains unclear whether these models are performing complex reasoning, or if they are merely learning complex surface correlation patterns (Davis and Marcus, 2015; Marcus, 2018). This difficulty in measuring the progress in commonsense reasoning using downstream tasks has yielded increased efforts at developing robust benchmarks for directly measuring commonsense capabilities in multiple settings, such as social interactions (Sap et al., 2019b; Rashkin et al., 2018a) and physical situations (Zellers et al., 2019; Talmor et al., 2019).

We hope that in the future, machines develop the kind of intelligence required to, for example, properly assist humans in everyday situations (e.g., a chatbot that anticipates the needs of an elderly person; Pollack, 2005). Current methods, however, are still not powerful or robust enough to be deployed in open-domain production settings, despite the clear improvements provided by large-scale pretrained language models. This shortcoming is partially due to inadequacy in acquiring, understanding and reasoning about commonsense knowledge, topics which remain understudied by the larger NLP, AI, and Vision communities relative to its importance in building AI agents. We organize this tutorial to provide researchers with information about the critical foundations and recent advances in commonsense, in the hopes of casting a brighter light on this promising area of future research.

In our tutorial, we will (1) outline the various types of commonsense (e.g., physical, social), and (2) discuss techniques to gather and represent commonsense knowledge, while highlighting the challenges specific to this type of knowledge (e.g., reporting bias). We will also (3) discuss the types of commonsense knowledge captured by modern NLP systems (e.g., large pretrained language models), (4) review ways to incorporate commonsense knowledge into downstream task models, and (5) present various benchmarks used to measure systems' commonsense reasoning abilities.

2 Description

What is commonsense? The tutorial will start with a brief overview of what commonsense is, how it is defined in the literature, and how hu-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 27–33
July 5, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

mans acquire it (Moore, 2013; Baron-Cohen et al., 1985). We will discuss notions of social commonsense (Burke, 1969; Goldman, 2015) and physical commonsense (Hayes, 1978; McRae et al., 2005). We will cover the differences between taxonomic and inferential knowledge (Davis and Marcus, 2015; Pearl and Mackenzie, 2018), and differentiate commonsense knowledge from related concepts (e.g., script learning; Schank and Abelson, 1975; Chambers and Jurafsky, 2008).

How to represent commonsense? We will review existing methods for representing commonsense, most of which focus solely on English. At first, symbolic logic approaches were the main representation type (Forbus, 1989; Lenat, 1995). While still in use today (Davis, 2017; Gordon and Hobbs, 2017), computational advances have allowed for more data-driven knowledge collection and representation (e.g., automatic extraction; Etzioni et al., 2008; Zhang et al., 2016; Elazar et al., 2019). We will cover recent approaches that use natural language to represent commonsense (Speer et al., 2017; Sap et al., 2019a), and while noting the challenges that come with using data-driven methods (Gordon and Van Durme, 2013; Jastrzebski et al., 2018).

What do machines know? Pretrained language models (LMs) have recently been described as "rediscovering the NLP pipeline" (Tenney et al., 2019a), i.e. replacing previous dedicated components of the traditional NLP pipeline, starting from low- and mid-level syntactic and semantic tasks (POS tagging, parsing, verb agreement, e.g., Peters et al., 2018; Jawahar et al., 2019; Shwartz and Dagan, 2019, *inter alia*), to high-level semantic tasks such as named entity recognition, coreference resolution and semantic role labeling (Tenney et al., 2019b; Liu et al., 2019a). We will discuss recent investigations into pretrained LMs' ability to capture world knowledge (Petroni et al., 2019; Logan et al., 2019) and learn or reason about commonsense (Feldman et al., 2019).

How to incorporate commonsense knowledge into downstream models? Given that large number of NLP applications are designed to require commonsense reasoning, we will review efforts to integrate such knowledge into NLP tasks. Various works have looked at directly encoding commonsense knowledge from structured KBs as additional inputs to a neural network in generation (Guan et al., 2018), dialogue (Zhou et al., 2018), QA (Mihaylov and Frank, 2018; Bauer et al., 2018; Lin et al., 2019; Weissenborn et al., 2017; Musa et al., 2019), and classification (Chen et al., 2018; Paul and Frank, 2019; Wang et al., 2018) tasks. For applications without available structured knowledge bases, researchers have relied on commonsense aggregated from corpus statistics pulled from unstructured text (Tandon et al., 2018; Lin et al., 2017; Li et al., 2018; Banerjee et al., 2019). More recently, rather than providing relevant commonsense as an additional input to neural networks, researchers have looked into indirectly encoding commonsense knowledge into the parameters of neural networks through pretraining on commonsense knowledge bases (Zhong et al., 2018) or explanations (Rajani et al., 2019), or by using multi-task objectives with commonsense relation prediction (Xia et al., 2019).

How to measure machines' ability of commonsense reasoning? We will explain that, despite their design, many natural language understanding (NLU) tasks hardly require machines to reason about commonsense (Lo Bue and Yates, 2011; Schwartz et al., 2017). This prompted efforts in creating benchmarks carefully designed to be impossible to solve without commonsense knowledge (Roemmele et al., 2011; Levesque, 2011).

In response, recent work has focused on using crowdsourcing and automatic filtering to design large-scale benchmarks while maintaining negative examples that are adversarial to machines (Zellers et al., 2018). We will review recent benchmarks that have emerged to assess whether machines have acquired physical (e.g., Talmor et al., 2019; Zellers et al., 2019), social (e.g., Sap et al., 2019b), or temporal commonsense reasoning capabilities (e.g., Zhou et al., 2019), as well as benchmarks that combine commonsense abilities with other tasks (e.g., reading comprehension; Ostermann et al., 2018; Zhang et al., 2018; Huang et al., 2019).

3 Outline

3.1 Schedule

Talk 1 (15 min.) will introduce and motivate this tutorial and discuss long term vision for NLP commonsense research.

Talk 2 (20 min.) will focus on the question "Do pre-trained language models capture commonsense knowledge?" and review recent work that studies what such models already capture due to their pre-training, what they can be fine-tuned to capture, and what types of knowledge are not captured.

Talk 3 (20 min.) will discuss ways of defining and representing commonsense, covering established symbolic methods and recent efforts for natural language representations.

Talk 4 (20 min.) will discuss neural and symbolic models of commonsense reasoning, focusing on models based on external knowledge integration for downstream tasks.

If time permits, we will end the first half with an interactive session and a preview to the second half.

Break (30 min.)

Talk 5 (20 min.) will continue the discussion on neural and symbolic models of commonsense knowledge representation, focusing on COMET (Bosselut et al., 2019), a language model trained on commonsense knowledge graphs. We will present its utility in a zero-shot model for a downstream commonsense question answering task.

Talk 6 (25 min.) will focus on temporal commonsense: how to represent it, how to incorporate it into downstream models, and how to test it.

Talk 7 (20 min.) will discuss ways to assess machine commonsense abilities, and challenges in developing benchmarks for such evaluations.

Concluding discussion (10 min.) will summarize the remaining challenges of commonsense research, and wrap up the tutorial.

3.2 Breadth

Due to the research interests and output of the presenters, we estimate that approximately 30% of the tutorial will center around work done by the presenters (Rashkin et al., 2018b; Sap et al., 2019a; Bosselut et al., 2019; Rashkin et al., 2018a; Sap et al., 2019b; Zellers et al., 2018, 2019; Sakaguchi et al., 2019; Bosselut and Choi, 2019; Shwartz et al., 2020).

4 Prerequisites

We will not expect attendees to be familiar with previous research on commonsense knowledge representation and reasoning, but participants should be familiar with:

- Knowledge of machine learning and deep learning – recent neural network architectures (e.g., RNN, CNN, Transformers), as well as large pre-trained language models models (e.g., BERT, GPT, GPT2).
- Familiarity with natural language processing tasks – understanding the basic problem to solve in tasks such as question answering (QA), natural language generation (NLG), textual entailment/natural language inference (NLI), etc.

5 Reading List

- Storks et al. (2019) – a survey on commonsense
- Levesque (2011) – The Winograd Schema challenge, considered an ideal benchmark for evaluating commonsense reasoning
- Speer et al. (2017) – A description of a prototypical commonsense knowledge base, its structure, and its curation
- Gordon and Van Durme (2013) – Overview of issues surrounding reporting bias, making automatic commonsense acquisition difficult
- Mostafazadeh et al. (2016) – A dataset that appears often in recent commonsense research
- Talmor et al. (2019) – One approach for leveraging crowdsourcing to construct a commonsense evaluation benchmark

6 Instructor information

Maarten Sap is a PhD student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. His research focuses primarily on social applications of NLP, specifically on endowing machines with social intelligence, social commonsense, or theory of mind.

Vered Shwartz is a postdoctoral researcher at the Allen Institute for Artificial Intelligence (AI2) and the Paul G. Allen School of Computer Science & Engineering at the University of Washington, working on lexical semantics, multiword expressions, and commonsense reasoning. She co-organized the ACL 2018 Student Research Workshop, the SemEval 2018 shared task on hypernymy discovery, and the AAAI 2020 Workshop on Reasoning for Complex Question Answering, Special Edition on Commonsense Reasoning.

Antoine Bosselut is a PhD student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington and a student researcher at the Allen Institute for Artificial Intelligence (AI2). His research interests are in integrating commonsense knowledge and reasoning into downstream applications for text generation, summarization, and conversational dialogue. He organized the West Coast NLP (WeCNLP) in 2018 and 2019 and the NeuralGen workshop at NAACL 2019.

Yejin Choi is an associate professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington and also a senior research manager at AI2 overseeing the project Mosaic. Her research interests include language grounding with vision, physical and social commonsense knowledge, language generation with long-term coherence, conversational AI, and AI for social good. She was a recipient of Borg Early Career Award (BECA) in 2018, among the IEEEs AI Top 10 to Watch in 2015, a co-recipient of the Marr Prize at ICCV 2013, and a faculty advisor for the Sounding Board team that won the inaugural Alexa Prize Challenge in 2017. She was on the steering committee of the NeuralGen workshop at NAACL 2019.

Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, University of Pennsylvania, and a Fellow of the AAAS, the ACM, AAAI, and the ACL. In 2017 Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. He was the Editor-in-Chief of the Journal of Artificial Intelligence Research (JAIR) and a program co-chair of AAAI, ACL and CoNLL. Dan has presented several tutorials in conferences including at ACL, on entity linking, temporal reasoning, transferable representation learning, and more.

References

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *ACL*.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the Autistic Child have a "Theory of Mind"? *Cognition*, 21(1):37–46.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *EMNLP*.

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 *ACL*.

Kenneth Burke. 1969. *A grammar of motives*, volume 177. Univ of California Press.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, pages 789–797.

Jiaao Chen, Jianshu Chen, and Zhou Yu. 2018. Incorporating structured commonsense knowledge in story completion. In *AAAI*.

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

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 *NAACL-HLT*.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. In *IJCAI*.

Joshua Feldman, Joe Davison, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. In *EMNLP*.

Kenneth D. Forbus. 1989. Qualitative process theory.

Alvin I Goldman. 2015. *Theory of human action*. Princeton University Press.

Andrew S Gordon and Jerry R Hobbs. 2017. *A Formal Theory of Commonsense Psychology: How People Think People Think*. Cambridge University Press.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Extraction. In *Automated Knowledge Base Construction (AKBC) 2013: The 3rd Workshop on Knowledge Extraction, at CIKM*.

Jian Guan, Yansen Wang, and Minlie Huang. 2018. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI*.

David Gunning. 2018. Machine common sense concept paper.

Patrick J. Hayes. 1978. The naive physics manifesto.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP*, volume abs/1909.00277.

Stanislaw Jastrzebski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio, and Jackie Chi Kit Cheung. 2018. Commonsense mining as knowledge base completion? a study on the impact of novelty. In *Workshop on Generalization in the Age of Deep Learning at NAACL*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38:32–38.

Hector J. Levesque. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Qian Li, Ziwei Li, Jin-Mao Wei, Yanhui Gu, Adam Jatowt, and Zhenglu Yang. 2018. A multi-attention based neural network with external knowledge for story ending predicting task. In *COLING*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *ArXiv*, abs/1909.02151.

Hongyu Lin, Le Sun, and Xianpei Han. 2017. Reasoning with heterogeneous knowledge for commonsense machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *NAACL-HLT*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Peter Lo Bue and Alexander Yates. 2011. Types of Common-Sense knowledge needed for recognizing textual entailment. In *ACL*.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Gary Marcus. 2018. Deep learning: A critical appraisal.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tzvetan Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *ACL*.

Chris Moore. 2013. *The development of commonsense psychology*. Psychology Press.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Mark Johnson, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *HLT-NAACL*.

Ryan Musa, Xiaoyan Wang, Achille Fokoue, Nicholas Mattei, Maria Chang, Pavan Kapanipathi, Bassem Makni, Kartik Talamadupula, and Michael J. Witbrock. 2019. Answering science exam questions using query reformulation with background knowledge. In *AKBC 2019*.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. *ArXiv*, abs/1904.00676.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *EMNLP*.

Martha E. Pollack. 2005. Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment. *AI Magazine*, 26:9–24.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *ACL*.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling naive psychology of characters in simple commonsense stories. In *ACL*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018b. Event2mind: Commonsense inference on events, intents, and reactions. In *ACL*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SemEval@NAACL-HLT*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *EMNLP*.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans and knowledge. In *IJCAI*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *CoNLL*.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *arXiv cs.CL 2004.05483*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

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.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. 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 *EMNLP*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. In *International Conference on Learning Representations*.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael J. Witbrock. 2018. Improving natural language inference using external knowledge in the science questions domain. In *AAAI*.

Dirk Weissenborn, Tom'avs Kovcisk'y, and Chris Dyer. 2017. Dynamic integration of background knowledge in neural nlu systems. *CoRR*, abs/1706.02596.

Jiangnan Xia, Chenjie Wu, and Ming Yan. 2019. Incorporating relation knowledge into commonsense reading comprehension with multi-task learning. *ArXiv*, abs/1908.04530.

Wenhan Xiong, M. Y. Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete kbs with knowledge-aware reader. In *ACL*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *ACL*.

Shenmin 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*, abs/1810.12885.

Shenmin Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2016. Ordinal commonsense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2018. Improving question answering by commonsense-based pre-training. *ArXiv*, abs/1809.03568.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*.