How Commonsense Knowledge Helps with Natural Language Tasks: A Survey of Recent Resources and Methodologies

Yubo Xie and Pearl Pu
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{yubo.xie, pearl.pu}@epfl.ch

June, 2021

Abstract

In this paper, we give an overview of commonsense reasoning in natural language processing, which requires a deeper understanding of the contexts and usually involves inference over implicit external knowledge. We first review some popular commonsense knowledge bases and commonsense reasoning benchmarks, but give more emphasis on the methodologies, including recent approaches that aim at solving some general natural language problems that take advantage of external knowledge bases. Finally, we discuss some future directions in pushing the boundary of commonsense reasoning in natural language processing.

1 Introduction

Natural language processing involves a wide range of tasks, from those on the lexical and syntactical levels such as tokenization and part-of-speech tagging, to those on the semantic and pragmatic levels such as reference resolution and text generation. While the former type of tasks is relatively more straightforward and has been more-or-less successfully tackled by machines, higher-level semantic and pragmatic tasks require deeper understanding of natural languages and still remain challenging for machines, with some of them difficult even for humans. These tasks often require machines to reason over commonsense knowledge, something universally accepted by humans but usually implicitly stated. For example, if you see a six-foot-tall person holding a two-foot-tall person in his arms, and you are told they are father and son [Davis and Marcus, 2015], you immediately infer that the taller person should be the father. Though this seems easy and straightforward for humans, it is challenging for machines because the

answer cannot be merely inferred from the given context, but depends on some external commonsense knowledge (in this case, "normally a father should be taller than a baby son").

Since 1960s, numerous attempts have been made at endowing natural language processing systems with the ability of commonsense reasoning. It is believed that Bar-Hillel [1960] was the first to mention the importance of incorporating commonsense knowledge into natural language processing systems, in the context of machine translation. Following that, many natural language processing tasks have been proposed with the aim of assessing the intelligent systems' ability of conducting various types of commonsense reasoning. Broadly speaking, these tasks, a.k.a. benchmarks, mainly deal with two types of commonsense knowledge. One type of commonsense knowledge that humans generally possess is "naive physics", which involves inference of how physical objects interact with each other. For example, if one is told that a glass of water falls onto the floor, he/she will most likely infer that the glass shatters and the floor becomes wet. Another type of commonsense knowledge that humans have is "intuitive psychology". This type of knowledge enables us to infer people's behaviors, intents, or emotions. For example, one could easily tell that a person who has lost his/her job probably feels upset.

Perhaps the most intuitive way, or at least the first step, to enable machines to reason with commonsense knowledge is to manually create a knowledge base that ideally includes humans' commonsense knowledge from all aspects. However, given that it is almost impossible to bag all knowledge that humans have ever possessed, and it is often difficult and debating to decide which knowledge is considered common and should be included, this task is quite challenging. Nevertheless, many efforts have been taken on building commonsense knowledge bases focused on certain domains, in the hope of achieving the ultimate goal of encoding the complete set of humans' commonsense knowledge.

To have an understanding of the current status of commonsense reasoning in natural language processing, we present a survey of the popular commonsense knowledge bases in this paper, and then discusses various benchmarks that assess machines' ability of commonsense reasoning. We also put much emphasis on methodological papers that target at more general natural language processing tasks by taking advantage of external commonsense knowledge bases. Finally, we give some possible future directions in the hope of extending the boundary of commonsense reasoning in natural language processing.

2 Commonsense Knowledge Bases

In this section, we give an overview of some well-known commonsense knowledge bases. Different from the lexical resources (such as WordNet [Miller, 1995]) and factual knowledge bases (such as Wikidata[1]) widely used in natural language processing tasks, commonsense knowledge bases encode knowledge that is usually implicitly stated and considered obvious to most humans. In this paper, we

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

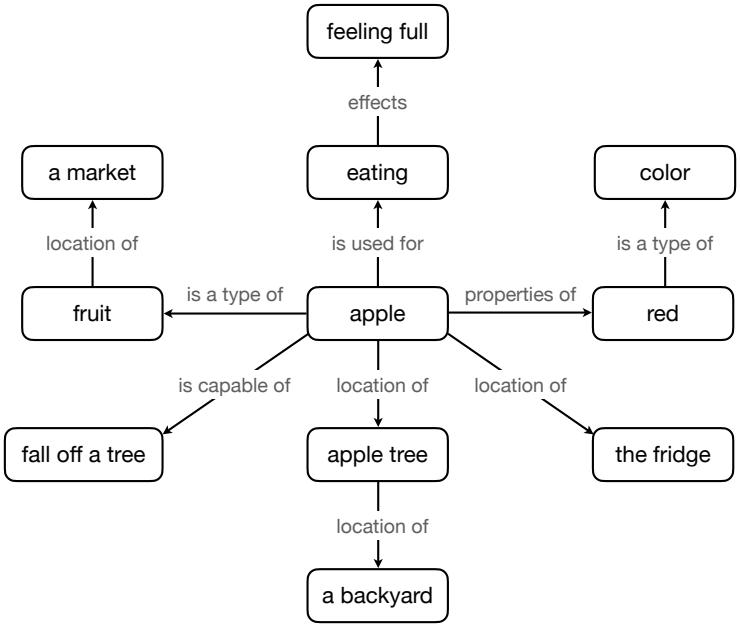

Figure 1: Entities and relations extracted from ConceptNet [Speer et al., 2017] that are related to the concept of *apple*.

do not pay much attention to the former two, but focus mainly on commonsense knowledge bases.

2.1 Knowledge Bases

Most commonsense knowledge bases are represented in the form of taxonomies, i.e., the knowledge base is usually a collection of individuals grouped into categories, with certain relations between them. A typical example is shown in Figure 1, which illustrates some entities and relations extracted from ConceptNet [Speer et al., 2017]. To our knowledge, most of the existing commonsense knowledge bases have a mixture of factual knowledge and commonsense knowledge. Table 1 gives a comparison of some well-known commonsense knowledge bases. Next, we are going to briefly introduce these knowledge bases one by one and then take a look at how these knowledge bases are evaluated. For each of the knowledge base, we also denote the year when its development began and the year when its newest version was published.

Cyc (1984–2012) Cyc [Lenat and Guha, 1989] is an artificial intelligence project aiming at integrating ontologies and commonsense knowledge from all different domains into one knowledge base, and based on that, achieving the ability of knowledge inference like human beings. Concepts in Cyc are called

Name	Size	Creation	Reference
OpenCyc	239,000 concepts 2,039,000 facts	Manual	Lenat and Guha [1989]
ConceptNet	8 million nodes 21 million links	Crowdsourcing	Speer et al. [2017]
SenticNet	200,000 concepts	Automatic	Cambria et al. [2020]
WebChild	2 million concepts 18 million assertions	Automatic	Tandon et al. [2017]
ATOMIC	309,515 nodes 877,108 triples	Crowdsourcing	Sap et al. [2019a]
ASER	194,000,677 nodes 64,351,959 relations	Automatic	Zhang et al. [2020]
CSKG	2,160,968 nodes 6,001,531 edges	Automatic	Ilievski et al. [2021]

Table 1: Comparison of some well-known commonsense knowledge bases.

"constants" and categorized into *individuals*, *collections*, *truth functions*, and *functions*. The Cyc project also includes an inference engine, which is capable of performing general logical deduction. Currently there are two releases of Cyc. OpenCyc 4.0 is the most recent public version and contains 239,000 concepts and 2,039,000 facts. ResearchCyc is licensed for research purposes and contains 500,000 concepts and 5,000,000 facts.

ConceptNet (1999–2017) ConceptNet [Speer et al., 2017] is a semantic network created by the Open Mind Common Sense (OMCS) [Singh et al., 2002], an artificial intelligence project aiming at building a large-scale commonsense knowledge base from the contributions of online users. It is a directed graph whose nodes are concepts, and the edges represent assertions of commonsense about the concepts, e.g., "is a", "is used for", "motivated by goal", etc. See Figure 1 for an example. The nodes are natural language phrases, e.g., noun phrases, verb phrases, or clauses. The latest version of the knowledge base is ConceptNet 5.5, which contains over 8 million nodes and over 21 million links.

SenticNet (2009–2020) As a knowledge base, SenticNet [Cambria et al., 2020] provides a set of semantics, sentics, and polarity associated with 200,000 natural language concepts. Specifically, semantics define the denotative information associated with natural language phrases, sentics define the emotion categorization values (expressed in terms of four affective dimensions) associated with these concepts, and polarity is floating number between -1 and $+1$. The knowledge base is automatically created from multiple other resources, e.g., WordNet-Affect [Strapparava and Valitutti, 2004] and OMCS [Singh et al., 2002].

WebChild (2014–2017) WebChild [Tandon et al., 2017] is a large-scale commonsense knowledge base that was automatically extracted and disambiguated from Web contents, using semi-supervised label propagation over graphs of noisy candidate assertions. The knowledge base contains triples that connect nouns with adjectives via fine-grained relations like "hasShape", "hasTaste", "evokesEmotion", etc. The arguments of these assertions, nouns and adjectives, are disambiguated by mapping them onto their proper WordNet senses. The newest version WebChild 2.0 was released in 2017 and contains over 2 million concepts with 18 million assertions about them.

ATOMIC (2019) ATOMIC [Sap et al., 2019a] is a commonsense knowledge graph consisting of 877K textual descriptions of inferential knowledge obtained from crowdsourcing. The knowledge graph focuses on *if-then* relations between events and possible inferences over the events. More specifically, the three types of relation are: "If-Event-Then-Mental-State", "If-Event-Then-Event", and "If-Event-Then-Persona". The base events were extracted from a variety of corpora including stories and books. In total the ATOMIC knowledge graph contains 309,515 nodes and 877,108 If-Event-Then-* triples.

ASER (2020) ASER [Zhang et al., 2020] is a large-scale eventuality knowledge graph automatically extracted from more than 11-billion-token unstructured textual data. It contains 15 relation types belonging to five categories, 194 million unique eventualities, and 64 million edges between them. The eventualities were extracted from a wide range of corpora from different sources, according to a selected set eventuality patterns. The eventuality relations were also automatically extracted using a selected set of seed connectives.

CSKG (2021) CommonSense Knowledge Graph (CSKG) [Ilievski et al., 2021] is a commonsense knowledge base created by consolidating and integrating several other key sources, using the proposed five principles. The resulted knowledge graph contains around 2 million nodes and 6 million edges between them.

2.2 Evaluation of Knowledge Bases

Since the application of commonsense knowledge bases is largely driven by their downstream tasks, there is no such universal and standard way to evaluate the quality of a commonsense knowledge base. Nevertheless, existing papers still adopt various ways to analyze the quality and scalability of the knowledge bases, and devise multiple experiments of downstream tasks to validate the effectiveness of the knowledge bases.

Comparison between knowledge bases Despite the fact that most of the existing commonsense knowledge bases come from different domains and possibly have different types of nodes (e.g., words, phrases, or events), several attempts have been made at mapping from one knowledge base to another and

ATOMIC Dimension	ConceptNet Relation
Wants	MotivatedByGoal, HasSubevent, HasFirstSubevent, CausesDesire
Effects	Causes, HasSubevent, HasFirstSubevent, HasLastSubevent
Needs	MotivatedByGoal, Entails, HasPrerequisite
Intents	MotivatedByGoal, CausesDesire, HasSubevent, HasFirstSubevent
Reactions	Causes, HasLastSubevent, HasSubevent, HasFirstSubevent
Attributes	HasProperty

Table 2: A mapping between ATOMIC and ConceptNet.

thus enabling the comparison between them. For example, the ATOMIC [Sap et al., 2019a] paper makes an attempt at comparing their commonsense knowledge base with ConceptNet [Speer et al., 2017]. They empirically investigated the difference between ATOMIC and ConceptNet, by mapping their dimensions and relations, which is shown in Table 2. The purpose of mapping and then comparing two knowledge bases is usually to check the coverage of the commonsense knowledge.

Downstream experiments To validate the effectiveness of commonsense knowledge bases, it is often helpful to run some downstream experiments by incorporating the knowledge base as an external resource. The ATOMIC [Sap et al., 2019a] paper investigated if a model could infer the *If-Then* commonsense given the previous event, by training a sequence-to-sequence model on the *If-Then* triples in the knowledge graph. The ASER [Zhang et al., 2020] paper proposed two ways to leverage the knowledge in ASER for solving the Winograd Scheme Challenge [Morgenstern et al., 2016]: string match and inference, and fine-tuning pre-trained language models. The CSKG [Ilievski et al., 2021] paper measured the relevance of CSKG for commonsense question answering tasks, as well as other popular natural language inferance tasks using pre-trained language models.

3 Benchmarks and Methodologies

Higher-level natural language processing tasks such as machine comprehension and natural language inference are challenging in that they often require machines to "think" like humans and infer over commonsense knowledge. Though

the goals of these proposed benchmarks and tasks vary from each other, successfully tackling them often demands the systems to have a certain degree of commonsense reasoning. In this section, we first review some of the popular natural language processing benchmarks that are at the semantic and pragmatic levels, which can be used to assess the systems' ability of performing commonsense knowledge inference. After that, we take a look at some methodologies that target at more general natural language problems (such as dialog generation) by utilizing external commonsense knowledge bases.

3.1 Benchmarks

In this section, we give a brief description of each benchmark according to their chronological order.

COPA The Choice of Plausible Alternatives (COPA) [Roemmele et al., 2011] involves causal inference between events. The dataset contains 1,000 examples in total, and in each example, an event is given, followed by a question asking for the correct effect or cause from two options.

Triangle-COPA Triangle-COPA [Gordon, 2016] is a variation of COPA [Roemmele et al., 2011] with 100 examples in the same format but accompanied with videos. The videos show situations where a circle and a triangle interact with each other. The questions asked are more focused on emotions and intentions.

ROCStories ROCStories [Mostafazadeh et al., 2016] contains 50,000 daily stories consisting of five sentences. During evaluation, a story is given, followed by two choices, of which one is a plausible ending and the other is an implausible ending. The system needs to choose the correct one from the two options.

Story Commonsense Story Commonsense [Rashkin et al., 2018a] is a dataset derived from ROCStories [Mostafazadeh et al., 2016] by annotating the emotions and motivations of the characters. The dataset contains 160,000 examples. Three classification tasks are involved, namely, the prediction of Maslow's basic human needs [Maslow, 1943], the prediction of Reiss' human motives [Reiss, 2004], and the prediction of Plutchik's eight emotions [Plutchik, 1980].

Event2Mind Event2Mind [Rashkin et al., 2018b] contains around 57,000 annotations of intents and reactions for around 25,000 events extracted from various corpora including ROCStories [Mostafazadeh et al., 2016]. Given an event with one or two participants, the system is supposed to predict the intents and reactions of the primary participant, and the reactions of the other participant.

SWAG Situations with Adversarial Generations (SWAG) [Zellers et al., 2018] is a dataset with 113,000 examples, where each example has a beginning sentence followed by four different endings. The system is supposed to choose the

most plausible ending from the four choices. The examples of the dataset were filtered adversarially to ensure the problem cannot be solved by simple and straightforward approaches.

SocialIQA SocialIQA [Sap et al., 2019b] is a crowdsourced benchmark containing 38,000 multiple choice questions for the purpose of probing emotional and social intelligence in everyday situations. Each example has a brief context and a question regarding the context, and the system is supposed to choose the correct answer from three options.

AlphaNLI Similar to ROCStories [Mostafazadeh et al., 2016], AlphaNLI [Bhagavatula et al., 2020] gives two observations as input, and the system needs to choose the most plausible hypothesized event from two choices, which is supposed to have happened between the two observations. The dataset contains around 170,000 examples.

3.1.1 Multi-task benchmarks

There are also benchmarks that contain multiple tasks and the performance of the participating system is usually measured by averaging its scores on each individual tasks. The GLUE benchmark [Wang et al., 2019b] is a popular collection of resources for the evaluation of natural language processing systems, which contains multiple tasks that would require systems to have the ability of commonsense reasoning, e.g., natural language inference, textual entailment, and question answering. The SuperGLUE benchmark [Wang et al., 2019a] improves the GLUE benchmark by incorporating more difficult natural language understanding tasks, to accommodate the increasingly powerful natural language processing systems in recent years.

3.2 Methodologies

As a general tool and resource, commonsense knowledge bases can be used in many natural language processing contexts, assisting in solving the tasks by enabling the systems to perform commonsense reasoning (at least to some extent). In this section, we take a look at some general natural language processing problems and see how external commonsense knowledge bases can be integrated into the development of various models.

3.2.1 Knowledge-based Question Answering

Question answering is a long-standing task in natural language processing where the goal is to answer natural language questions correctly, and therefore can be used to assess machines' understanding of some certain domain. This is also an area where knowledge bases can play an important role. Generally speaking, there are two types of question answering tasks in natural language processing.

One is *question answering on knowledge base*, which aims at answering the question using related information in the knowledge base. Most existing models [Xu et al., 2016, Dong et al., 2017, Hao et al., 2017] find related knowledge base entities in the given question using traditional rule-based methods. Lan et al. [2019] proposed to do entity linking using a generation-and-scoring approach to gradually refining the set of topic units, so that a wider range of knowledge base units could be covered. Another type of question answering is *question answering on text* (also called machine reading comprehension based question answering), where the goal is to answer the question based on a given passage of relevant content. Several attempts have been made at training end-to-end neural networks on large-scale question answering datasets such as SQuAD [Rajpurkar et al., 2016]. Bauer et al. [2018] approached the problem by proposing a model that selects grounded multi-hop relational commonsense information from ConceptNet [Speer et al., 2017] via a pointwise mutual information and term-frequency based scoring function. There is also work aiming at solving the task of question answering that is provided with both a knowledge base and some relevant text [Das et al., 2017].

3.2.2 Knowledge-Based Chatbots

The integration of commonsense knowledge bases into the development of chatbots aims at improving the diversity of the generated/selected response, by encoding extra knowledge information into the training process. Generally speaking, according to how the chatbots are implemented, we have the following categories.

Template-based Han et al. [2015] integrated Freebase [Bollacker et al., 2008] into a template-based dialog system that consists of five modules. Given a user utterance, the system extracts an important name entity from it, and then scans the knowledge base to extract contents related to this entity. The extracted contents are then used to fill in the slots in the response templates.

Retrieval-based Young et al. [2018] incorporated ConceptNet [Speer et al., 2017] into a retrieval-based chatbot. The model recovers the concepts in the input message using simple n-gram matching, and then encodes the assertions using an LSTM. The assertion with the highest score is then incorporated into the Tri-LSTM to calculate the final score of the candidate reply. The model was trained on the Twitter Dialogue Dataset, and compared with several baselines such as supervised word embeddings and memory networks [Bordes et al., 2017].

Generation-based Generation-based dialog systems often adopt an encoder-decoder model architecture, and the decoder is responsible for generating the response one token at a step. Ghazvininejad et al. [2018] generalized the widely used sequence-to-sequence approach by conditioning response generation on both the input conversation history and the external knowledge extracted from

the unstructured grounded dataset. Zhou et al. [2018] devised a static graph attention mechanism on the relevant knowledge subgraph retrieved from ConceptNet. The retrieved knowledge graph is used in the knowledge interpreter of the encoder as well as the knowledge aware generator in the decoder. Wu et al. [2020] proposed a felicitous fact recognizer that retrieves knowledge facts from the knowledge base by considering the specific dialog context.

4 Future Directions

Commonsense reasoning still remains a challenging problem in natural language processing. The amount of commonsense knowledge that humans possess is massive and thus a complete integration is difficult. Moreover, commonsense knowledge from different domains varies vastly from each other and some of them are only partially understood. Creating a commonsense knowledge base manually or via crowdsourcing is often quite time and money consuming, while automatic extraction can suffer from the problem of accuracy. Nevertheless, given the current status of the community of commonsense reasoning in natural language processing, we think there are still some future directions that could bring interesting research insights.

Social intelligence in conversations While some of the existing commonsense knowledge bases [Speer et al., 2017, Sap et al., 2019a] do contain knowledge of intuitive psychology of human beings (e.g., emotion states, intents, possible behaviors), they rarely emphasize on the social intelligence found in daily human-human conversations. This includes the social interaction patterns that are widely adopted by human beings, e.g., how to respond to others' shared experience in an empathetic way. Apparently this common knowledge of social intelligence is embedded in people's daily conversations and can be extracted from the abundant dialog resources on the Web. We expect that the construction of such social intelligence knowledge bases could be beneficial to chatbots that are supposed to generate responses that are more diverse or empathetic.

Commonsense knowledge in text classification Some text classification systems [Yamada and Shindo, 2019, Chen et al., 2019] use external knowledge bases (for example entities extracted from Wikipedia pages) to facilitate the prediction of the text category. Usually these text classification tasks focus on topics or sentiments, and the involved knowledge is often on the lexical-level, or just facts of the world. It can be imagined that for tasks that require more subtle natural language inference, e.g., humor or sarcasm recognition, some commonsense knowledge could be useful. Therefore, one possible direction in the context of these text classification tasks is to considering using knowledge bases more dedicated to intuitive psychology.

Automatic construction of knowledge bases With pre-trained language models such as BERT [Devlin et al., 2019] and GPT [Radford et al., 2019] being

popular in recent years, there have been attempts at automatically proposing new entities and relations for an existing knowledge base [Bosselut et al., 2019]. Though the existing work shows the feasibility of applying pre-trained language models for constructing knowledge bases automatically, the results were only obtained on small test sets, and there has not been a large-scale commonsense knowledge base that is created by utilizing the knowledge learned in pre-trained language models and has good quality at the same time.

5 Conclusion

This paper gives a survey of some popular commonsense knowledge bases, and then discusses some benchmarks and tasks that can be used to assess natural language processing systems' ability of commonsense reasoning. We also review some of the methodological papers that target at more general natural language processing tasks by taking advantage of external commonsense knowledge bases. Finally, we give some possible future directions in the hope of pushing the boundary of commonsense reasoning in natural language processing.

References

Yehoshua Bar-Hillel. The present status of automatic translation of languages. *Adv. Comput.*, 1:91–163, 1960. doi: 10.1016/S0065-2458(08)60607-5. URL https://doi.org/10.1016/S0065-2458(08)60607-5.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. Commonsense for generative multi-hop question answering tasks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of EMNLP 2018*, pages 4220–4230. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1454. URL https://doi.org/10.18653/v1/d18-1454.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *Proceedings of ICLR 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In Jason Tsong-Li Wang, editor, *Proceedings of SIGMOD 2008*, pages 1247–1250. ACM, 2008. doi: 10.1145/1376616.1376746. URL https://doi.org/10.1145/1376616.1376746.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In *Conference Track Proceedings of ICLR 2017*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1Bb3D5gg.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic

knowledge graph construction. In *Proceedings of ACL 2019*, pages 4762–4779. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1470. URL https://doi.org/10.18653/v1/p19-1470.

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of CIKM 2020*, pages 105–114. ACM, 2020. doi: 10.1145/3340531.3412003. URL https://doi.org/10.1145/3340531.3412003.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. Deep short text classification with knowledge powered attention. In *Proceedings of AAAI 2019*, pages 6252–6259. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33016252. URL https://doi.org/10.1609/aaai.v33i01.33016252.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of ACL 2017*, pages 358–365. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-2057. URL https://doi.org/10.18653/v1/P17-2057.

Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, 2015. doi: 10.1145/2701413. URL https://doi.org/10.1145/2701413.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of EMNLP 2017*, pages 875–886. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1091. URL https://doi.org/10.18653/v1/d17-1091.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *Proceedings of AAAI 2018*, pages 5110–5117. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16710.

Andrew S. Gordon. Commonsense interpretation of triangle behavior. In *Proceedings of AAAI 2016*, pages 3719–3725. AAAI Press, 2016. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11790.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of SIGDIAL 2015*, pages 129–133. The Association for Computer Linguistics, 2015. doi: 10.18653/v1/w15-4616. URL https://doi.org/10.18653/v1/w15-4616.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of ACL 2017*, pages 221–231. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1021. URL https://doi.org/10.18653/v1/P17-1021.

Filip Ilievski, Pedro A. Szekely, and Bin Zhang. CSKG: the commonsense knowledge graph. In *Proceedings of ESWC 2021*, volume 12731 of *Lecture Notes in Computer Science*, pages 680–696. Springer, 2021. doi: 10.1007/978-3-030-77385-4_41. URL https://doi.org/10.1007/978-3-030-77385-4_41.

Yunshi Lan, Shuohang Wang, and Jing Jiang. Knowledge base question answering with topic units. In Sarit Kraus, editor, *Proceedings of IJCAI 2019*, pages 5046–5052. ijcai.org, 2019. doi: 10.24963/ijcai.2019/701. URL https://doi.org/10.24963/ijcai.2019/701.

Douglas B Lenat and Ramanathan V Guha. *Building large knowledge-based systems: Representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

Abraham Harold Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

George A. Miller. Wordnet: A lexical database for English. *Commun. ACM*, 38 (11):39–41, 1995. doi: 10.1145/219717.219748. URL http://doi.acm.org/10.1145/219717.219748.

Leora Morgenstern, Ernest Davis, and Charles L. Ortiz Jr. Planning, executing, and evaluating the winograd schema challenge. *AI Mag.*, 37(1):50–54, 2016. doi: 10.1609/aimag.v37i1.2639. URL https://doi.org/10.1609/aimag.v37i1.2639.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT 2016*, pages 839–849. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1098. URL https://doi.org/10.18653/v1/n16-1098.

Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1264. URL https://doi.org/10.18653/v1/d16-1264.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings ACL 2018*, pages 2289–2299. Association for Computational Linguistics, 2018a. doi: 10.18653/v1/P18-1213. URL https://www.aclweb.org/anthology/P18-1213/.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings ACL 2018*, pages 463–473. Association for Computational Linguistics, 2018b. doi: 10.18653/v1/P18-1043. URL https://www.aclweb.org/anthology/P18-1043/.

Steven Reiss. Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of general psychology*, 8(3):179–193, 2004.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06*. AAAI, 2011. URL http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2418.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of AAAI 2019*, pages 3027–3035. AAAI Press, 2019a. doi: 10.1609/aaai.v33i01.33013027. URL https://doi.org/10.1609/aaai.v33i01.33013027.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of EMNLP-IJCNLP 2019*, pages 4462–4472. Association for Computational Linguistics, 2019b. doi: 10.18653/v1/D19-1454. URL https://doi.org/10.18653/v1/D19-1454.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open Mind Common Sense: Knowledge acquisition from the general public. In Robert Meersman and Zahir Tari, editors, *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36124-4.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI 2017*, pages 4444–4451. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *Proceedings of LREC 2004*. European Language Resources Association, 2004. URL http://www.lrec-conf.org/proceedings/lrec2004/summaries/369.htm.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. WebChild 2.0 : Fine-grained commonsense knowledge distillation. In Mohit Bansal and Heng Ji, editors, *Proceedings of ACL 2017*, pages 115–120. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-4020. URL https://doi.org/10.18653/v1/P17-4020.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS 2019*, pages 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of ACL 2020*, pages 5811–5820. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.515. URL https://doi.org/10.18653/v1/2020.acl-main.515.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. Question answering on freebase via relation extraction and textual evidence. In *Proceedings of ACL 2016*. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/p16-1220. URL https://doi.org/10.18653/v1/p16-1220.

Ikuya Yamada and Hiroyuki Shindo. Neural attentive bag-of-entities model for text classification. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of CoNLL 2019*, pages 563–573. Association for Computational Linguistics, 2019. doi: 10.18653/v1/K19-1052. URL https://doi.org/10.18653/v1/K19-1052.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of AAAI 2018*, pages 4970–4977. AAAI Press,

2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16573.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of EMNLP 2018*, pages 93–104. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1009. URL https://doi.org/10.18653/v1/d18-1009.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. ASER: A large-scale eventuality knowledge graph. In *Proceedings of WWW 2020*, pages 201–211. ACM/IW3C2, 2020. doi: 10.1145/3366423.3380107. URL https://doi.org/10.1145/3366423.3380107.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of IJCAI 2018*, pages 4623–4629. ijcai.org, 2018. doi: 10.24963/ijcai.2018/643. URL https://doi.org/10.24963/ijcai.2018/643.