Commonsense Reasoning: Benchmarks and Resources

Jun Yan

University of Southern California
yanjun@usc.edu

November 19, 2019

1 Introduction

Commonsense knowledge is the basic level of practical knowledge. It concerns everyday situations and events in the world and is commonly shared among most people. For example, *it's ok to keep the closet door open, but it's not ok to keep the fridge door open* as the food inside might go bad. There is some exceptional cases like when there is no food in the refrigerator, but this statement is considered as commonsense because it's agreed by most people. Humans need commonsense knowledge to live and interact with each other in a reasonable and safe way, and Artificial Intelligence (AI) needs commonsense to understand humans better.

Especially, when humans use their languages to communicate with each other, they often rely on commonsense knowledge, which acts as implicit assumptions and makes human's languages concise without lacking precision. However, machines by nature don't have such background knowledge. Machine learning models can't accumulate human's commonsense knowledge through interacting with the environment. Therefore, empowering Natural Language Processing (NLP) techniques with knowledge is one of the major longterm goals for Artificial Intelligence.

In this presentation, I'll introduce benchmarks and resources for commonsense reasoning. The roles of benchmarks and resources in the commonsense-related research are depicted in Figure 1.

2 Benchmarks

Since 2005, there has been a surge of commonsense-directed benchmarks being created (see Figure 2). These benchmarks can be categorized according to the classic NLP problems they are based on. Note that while commonsense knowledge is usually implicit, there's another kind of knowledge called **common knowledge**, which refers to specific facts about the world that are often explicitly stated. A more clear way to distinguish commonsense knowledge and common knowledge should be: **Commonsense knowledge** is about finding wisdom from experience. It's sometimes wrong. **Common knowledge** is the generally known information. It is not contextual and is not questioned. For example, "getting wet in winter leads to cold" is commonsense knowledge, but "water is wet" is common knowledge. Usually both kinds of knowledge are needed during reasoning, so we don't put a strict boundary between them when introducing the benchmarks.

2.1 Coreference Resolution

Coreference resolution aims to determine which entity or event in a text a particular pronoun refers to.

Winograd Schema Challenge [1]

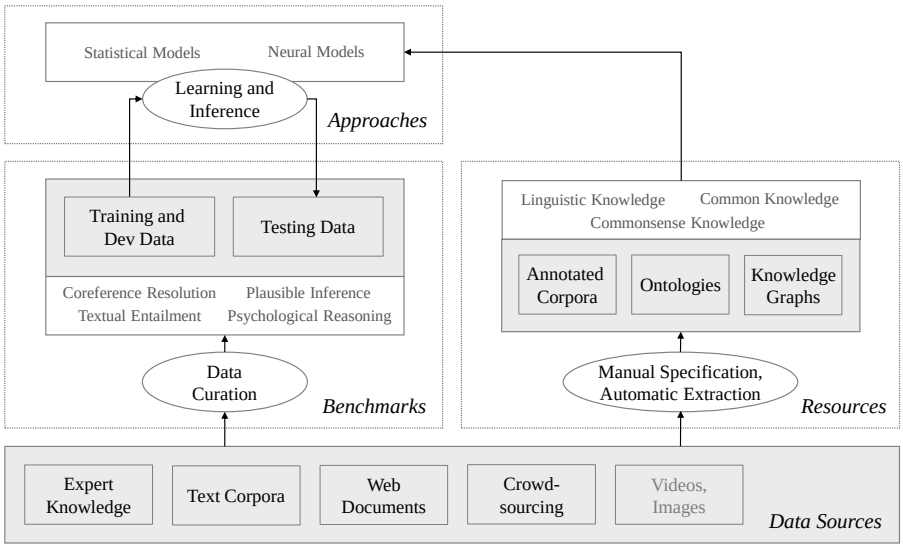

Figure 1: Main research efforts in commonsense knowledge and reasoning from the NLP community occur in three areas: benchmarks and tasks, knowledge resources, and learning and inference approaches.

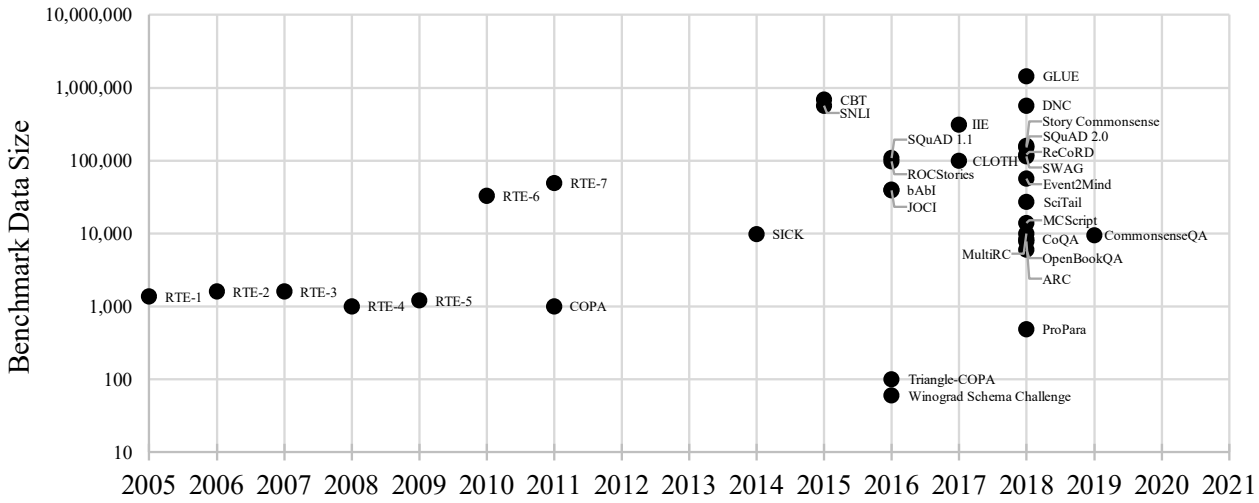

Figure 2: A surge of commonsense-directed benchmarks being created since 2005.

In this challenge, a system must disambiguate a pronoun whose coreferent may be one of two entities, and can be changed by replacing a single word in the sentence.

> The trophy would not fit in the brown suitcase because it was too big (small). What was too big (small)?
>
> (A) the trophy (B) the suitcase

It doesn't have train or dev set. The test set has 273 instances. The SOTA model is GPT-2, which achieves 70.70% accuracy.

2.2 Question Answering

ARC [2]

The AI2 Reasoning Challenge (ARC) provides a datasets of four-way multiple-choice science questions and answers, as well as a corpus of 14 million science-related sentences which are claimed to contain most of the information needed to answer the questions. The questions cover 8 reasoning types: Definition, Basic Facts & Properties, Structure, Processes & Causal, Teleology & Purpose, Algebraic, Experiments, Spatial / Kinematic.

> Which item below is not made from a material grown in nature?
>
> (A) a cotton shirt (B) a wooden chair (C) a plastic spoon (D) a grass basket

The challenge dataset contains 1,119/299/1,172 (tain/dev/test) instances. The SOTA model (FreeLB-RoBERTa) achieves 68% accuracy.

MCScript [3]

Scripts are sequences of events describing stereotypical human activities. MCScript is a dataset for assessing the contribution of script knowledge to machine comprehension.

> Text: I wanted to plant a tree. I went to the home and garden store and picked a nice oak. Afterwards, I planted it in my garden.
>
> Q1: What was used to dig the hole? (A) a shovel (B) his bare hands
>
> Q2: When did he plant the tree? (A) after watering it (B) after taking it home

The dataset consists of 9,731/1,411/2,797 instances based on short passages. The SOTA model (Muti-Perspective Fusion Network) achieves 84.84% accuracy.

ProPara [4]

ProPara (Process Paragraphs) is a dataset containing 488 human-authored paragraphs of procedural text, along with 81k annotations about the changing states (existence and location) of entities in those paragraphs. The end-task is to predict location and existence changes that occur.

> Chloroplasts in the leaf of the plant trap light from the sun. The roots absorb water and minerals from the soil. This combination of water and minerals flows from the stem into the leaf. Carbon dioxide enters the **leaf**. Light, water and minerals, and the carbon dioxide all combine into a mixture. This mixture forms **sugar** (glucose) which is what the plant eats.
>
> Q: Where is sugar produced? A: in the leaf

The SOTA model (NCET) achieves 62.50% F1 score[1] while human performance is 83.90%.

MultiRC [5]

MultiRC is a reading comprehension dataset consisting of about 10,000 questions posed on over 800 paragraphs across a variety of topic domains. Most questions can only be answered by reasoning over multiple sentences. Answers are not spans of text from the paragraph, and the number of answer choices as well as the number of correct answers for each question is variable. It includes a variety of nontrivial semantic phenomena in passages, such as coreference and causal relationships, which often require commonsense to recognize and parse.

The SOTA model (T5) achieves 88.20% F1 score while human performance is 81.80% (RoBERTa is 84.4%).

OpenBookQA [6]

[1]There can be multiple answers to a question like *"When is e moved?"*.

OpenBookQA contains about 6,000 4-way multiple choice science questions which may require science facts or other common and commonsense knowledge. It provides an "open book" of about 1,300 science facts to support answering the questions, each associated directly with the question(s).

> Question: Which of these would let the most heat travel through?
>
> (A) a new pair of jeans. (B) a steel spoon in a cafeteria.
>
> (C) a cotton candy at a store. (D) a calvin klein cotton hat.
>
> Science Fact (from “open book”): Metal is a thermal conductor.
>
> Common Knowledge: Steel is made of metal. Heat travels through a thermal conductor.

The SOTA model (AristoRoBERTaV7) achieves 77.80% accuracy, while human performance is 91.70%.

CommonsenseQA [7]

Commonsense QA consists of 9,500 five-way multiple-choice questions. To ensure an emphasis on commonsense, each question requires one to disambiguate a target concept from three connected concepts in ConceptNet, a commonsense knowledge graph.

> Where on a river can you hold a cup upright to catch water on a sunny day?
>
> (A) waterfall (B) bridge (C) valley (D) pebble (E) mountain

The SOTA model (XLNet + Graph Reasoning) achieves 75.30% accuracy, while human performance is 88.90%.

2.3 Textual Entialment

Textual entailment is defined as a directional relationship between a text and a hypothesis. It can be said that the text entails the hypothesis if a typical person would infer that the hypothesis is true given the text. A positive example could be: Text: *If you help the needy, God will reward you.* Hypothesis: *Giving money to a poor man has good consequences.*

Recognizing textual entailment (RTE, also known as Natural Language Inference, NLI) requires the utilization of several simpler language processing skills, such as paraphrase, object tracking, and causal reasoning, but since it also requires a sense of what a typical person would infer, commonsense knowledge is often essential to textual entailment tasks.

SNLI [8]

The Stanford Natural Language Inference (SNLI) benchmark provides a three-way decision task. It contains nearly 600,000 sentence pairs.

> Test: A man inspects the uniform of a figure in some East Asian country.
>
> Hypothesis: The man is sleeping.
>
> (A) entailment (B) contradiction (C) neutral

The SOTA model (MT-DNN) achieves 91.60% accuracy. The estimated human performance is 87.70%.

SciTail [9]

The SciTail dataset is created from multiple-choice science exams and web sentences. Each question and the correct answer choice are converted into an assertive statement to form the hypothesis. Information retrieval is used to obtain relevant text from a large text corpus of web sentences to form the premise (text). SciTail consists of about 27,000 premise-hypothesis sentence pairs adapted from

science questions into a 2-way (entails/neutral) entailment task. It is primarily science-based, which may require some knowledge more advanced than everyday commonsense.

> Text: Water and other materials necessary for biological activity in trees are transported throughout the stem and branches in thin, hollow tubes in the xylem, or wood tissue.
>
> Hypothesis: Stems transport water to other parts of the plant through a system of tubes.
>
> (A) entailment (B) neutral

The SOTA model (KD-MT-DNN) achieves 96.10% accuracy.

2.4 Multiple Tasks

bAbI [10]

The bAbI benchmark consists of 20 prerequisite tasks, each with 1,000 examples for training and 1,000 for testing. Each task presents systems with a passage, then asks a reading comprehension question, but each task focuses on a different type of reasoning or language processing task. Tasks are as follows:

> Single supporting fact; Two supporting facts; Three supporting facts; Two argument relations; Three argument relations; Yes/no questions; Counting; Lists/sets; Simple negation; Indefinite knowledge; Basic coreference; Conjunction; Compound coreference; Time reasoning; Basic deduction; Basic induction; Positional reasoning; Size reasoning; Path finding; Agent's motivations

Figure 3 shows some examples.

(A) Task 15: Basic Deduction
Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of?
wolves

(B) Task 16: Basic Induction
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg?
white

(C) Task 17: Positional Reasoning
The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red square to the left of the triangle?
yes

(D) Task 18: Size Reasoning
The football fits in the suitcase.
The suitcase fits in the cupboard.
The box is smaller than the football.
Will the box fit in the suitcase?
yes

Figure 3: Examples of commonsense reasoning tasks from bAbI. Answers in bold.

3 Knowledge Resources

The lack of this commonsense knowledge is one of the major bottlenecks in machine intelligence. In order to remove this bottleneck, decades of efforts have been made in developing various knowledge resources in the field of AI. To understand human language, it is important to have linguistic knowledge resources that allow computers to identify syntactic and semantic structures from language. These structures often need to be augmented with common knowledge and commonsense knowledge in order to reach a full understanding.

3.1 Linguistic Knowledge

WordNet [11]

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The graph is directed and acyclic (a DAG), though not necessarily a tree since each synset can have several hypernyms.

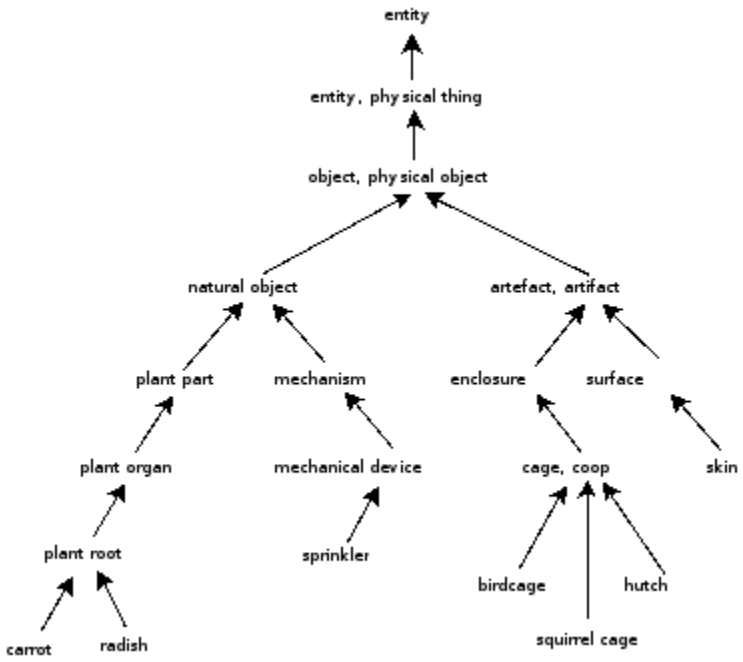

Figure 4: Examples of "is a" relation in WordNet.

WordNet contains 155,327 words organized in 175,979 synsets for a total of 207,016 word-sense pairs. The relations for nouns include hypernym, hyponym, coordinate term, meronym, and holonym. The relations for verbs include hypernym, troponym, entailment, and coordinate term. An example is presented in Figure 4.

3.2 Common Knowledge

DBPedia [12]

DBPedia is a Wikipedia-based knowledge base originally consisting of structured knowledge from more than 1.95 million Wikipedia articles. The English version of DBpedia knowledge base describes 4.58 million things, out of which 4.22 million are classified in a consistent ontology, including 1,445,000 persons, 735,000 places, 411,000 creative works, 241,000 organizations, 251,000 species and 6,000 diseases. An example is presented in Figure 5.

Freebase [13]

Freebase is a knowledge graph which originally contained 125 million triples of general human knowledge about 4,000 types of entities and 7,000 properties of entities. This resource was later absorbed into the Google Knowledge Graph. The latest release contains more than 1.9 billion triples.

3.3 Commonsense Knowledge

ConceptNet [14]

ConceptNet originated from the crowdsourcing project Open Mind Common Sense. It is a multilingual semantic network, designed to help computers understand the meanings of words that people

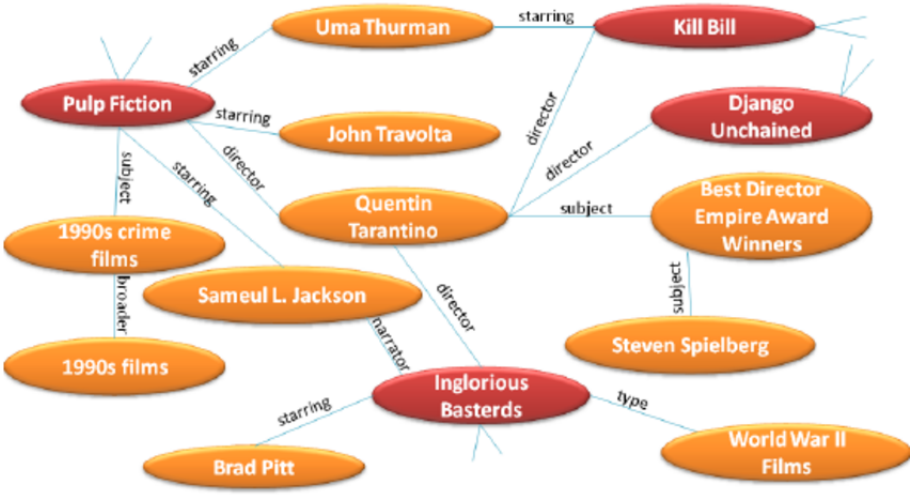

Figure 5: Example of a graph extracted from DBpedia.

use. ConceptNet 5.5 contains over 21 million links between over 8 million nodes. An example is presented in Figure 6.

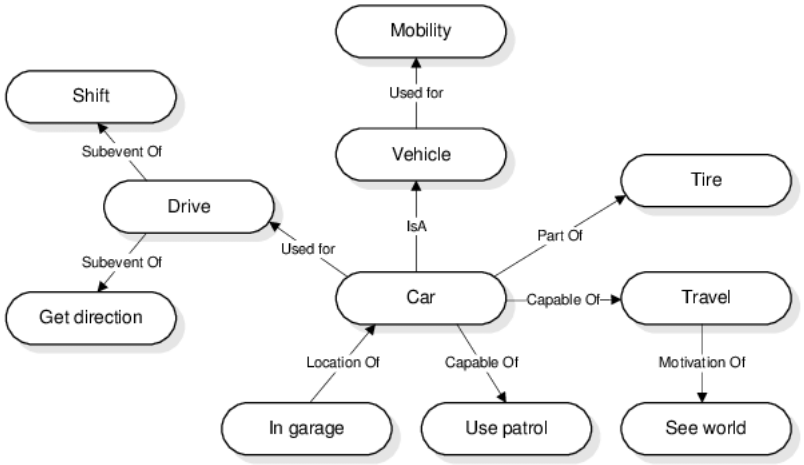

Figure 6: Concepts related to a car in ConceptNet.

ATOMIC [15]

The Atlas of Machine Commonsense (ATOMIC) is a knowledge graph consisting of about 300,000 nodes corresponding to short textual descriptions of events, and about 877,000 "if-event-then" triples representing if-then relationships between everyday events. An example is presented in Figure 7. The authors demonstrate that neural models can learn simple commonsense reasoning skills from ATOMIC to make inferences about previously unseen events. For example, given “PersonX bakes bread. As a result, X will”, the model can generate “get dirty”, “eat food”, which are not included in the original knowledge graph.

LocatedNear [16]

The authors claim that objects which tend to be near each other (e.g., silverware, a plate, and a glass) is a type of commonsense knowledge lacking in previous knowledge bases like ConceptNet 5.5, so they create two datasets describing LocatedNear property. The first consists of 5,000 sentences describing scenes of two objects labeled for whether the objects tend to occur near each other. The second consists of 500 pairs of objects with human-produced confidence scores for how likely the objects are to appear near each other.

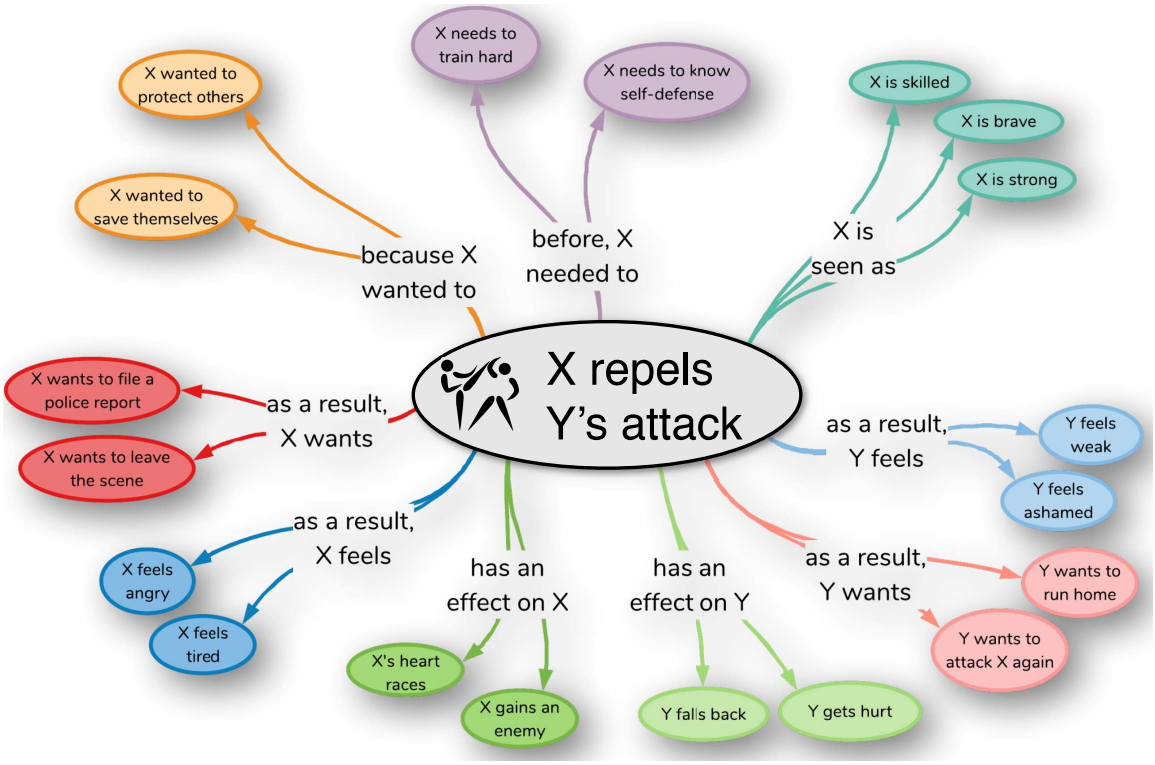

Figure 7: An example of facts in ATOMIC.

Acknowledgments

Most content of this presentation is heavily relied on [More1]. The authors provided a comprehensive survey of commonsense reasoning's benchmarks, resources and approaches. Please refer to it to get more details.

References

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

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[3] Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. Mcscript: A novel dataset for assessing machine comprehension using script knowledge. *arXiv preprint arXiv:1803.05223*, 2018.

[4] Bhavana Dalvi Mishra, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. *arXiv preprint arXiv:1805.06975*, 2018.

[5] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, 2018.

[6] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[7] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[8] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[9] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[11] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[12] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[13] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

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

[15] 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 the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019.

[16] Frank F Xu, Bill Yuchen Lin, and Kenny Q Zhu. Automatic extraction of commonsense locatednear knowledge. *arXiv preprint arXiv:1711.04204*, 2017.

Reading Materials

[More1] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.