Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches

Shane Storks   SSTORKS@UMICH.EDU
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109-2121 USA

Qiaozi Gao   GAOQIAOZ@MSU.EDU
Computer Science and Engineering
Michigan State University
East Lansing, MI 48824 USA

Joyce Y. Chai   CHAIJY@UMICH.EDU
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109-2121 USA

Abstract

In the NLP community, recent years have seen a surge of research activities that address machines' ability to perform deep language understanding which goes beyond what is explicitly stated in text, rather relying on reasoning and knowledge of the world. Many benchmark tasks and datasets have been created to support the development and evaluation of such natural language inference ability. As these benchmarks become instrumental and a driving force for the NLP research community, this paper aims to provide an overview of recent benchmarks, relevant knowledge resources, and state-of-the-art learning and inference approaches in order to support a better understanding of this growing field.

1. Introduction

We humans use a variety of knowledge and reasoning to help understand meanings of language. For example, consider these sentences from Minsky (2000): "Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound." From this, it is not difficult for us to understand that Jack did not find any money, and because of that, Jack was having a negative emotion. What made us come to this conclusion, which was not explicitly stated in the passage, is the knowledge we have about the world and the underlying reasoning process, often called *commonsense thought* (Minsky, 2000) or *commonsense reasoning* (Davis & Marcus, 2015), that allows us to connect pieces of knowledge to reach the new conclusion. We know that a *piggy bank* is a container that holds coins (not pigs) and that coins are money which are made of metal. Since metal is a hard solid, the coins will make a sound when shaken inside of a container such as a piggy bank; if there is no sound, then there are no coins. It is also likely that we can predict that as piggy banks are typically possessed by children, there is a good chance that Jack is a child. Alternatively, these predictions may be derived from similar events we have experienced as children, and allow us to make similar conclusions by analogy (Minsky, 2000). While this kind of knowledge and reasoning comes so naturally to human readers, it is notoriously difficult for machines. Despite
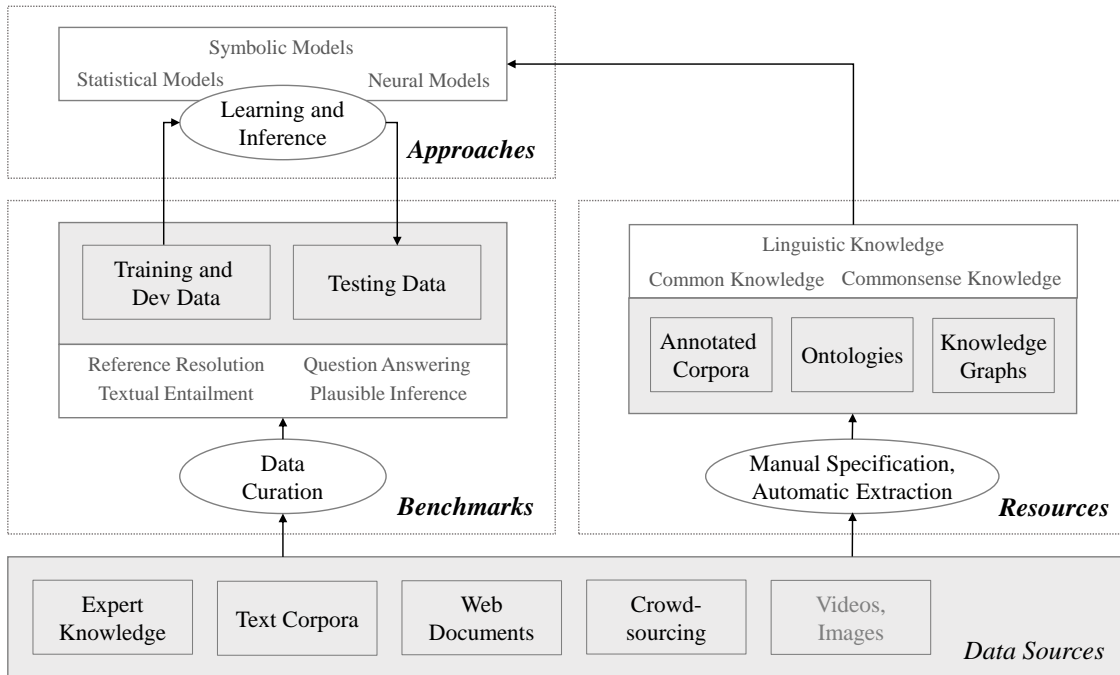
Figure 1: Main research efforts in natural language inference from the NLP community occur in three areas: benchmarks and tasks, knowledge resources, and learning and inference approaches.[2]

significant advances in natural language processing in the last several decades, machines are still far away from having this type of *natural language inference (NLI)* ability.

To address this problem, recent years have seen a surge of research activities on NLI: *machines' capability of deep understanding of language that goes beyond what is explicitly expressed, rather relying on new conclusions inferred from knowledge about how the world works.*[3] Many benchmark datasets have been created to help develop and evaluate NLI algorithms and models. These datasets have drawn significant attention from the research community, and many learning and inference approaches have been developed. To facilitate quantitative evaluation and encourage broader participation, various leaderboards for benchmarks have been set up and maintained. As these benchmarks become instrumental and a driving force for the community, this paper aims to provide an overview of recent advances in NLI focusing on existing tasks and benchmarks, knowledge resources, and learning and inference approaches, and discuss current limitations and future opportunities. We hope this paper will provide an entry point to those who are not familiar with but interested in pursuing research in this quickly evolving topic area.

---

2. Although videos and images are often used for creating various benchmarks, they are not the focus of this paper.

3. While NLI is often a term used to refer to textual entailment tasks, e.g., by Bowman, Angeli, Potts, and Manning (2015), we use the term here to refer to a broader sense of inference which includes but is not restricted to the textual entailment task formulation.

The NLP community has a long history of creating benchmarks to facilitate algorithm development and evaluation for various language processing tasks. Many resources to support those approaches to the benchmarks have also been proposed. It is neither possible nor intended for this paper to provide a comprehensive discussion of all benchmarks and resources proposed by the NLP community. This survey focuses on the ongoing research effort in benchmarks, resources, and approaches specifically toward natural language inference (NLI), as shown in Figure 1.

1.1 Scope of Benchmark Datasets

There has been a long line of work toward creating benchmark datasets to evaluate machines' progress on language processing tasks, and facilitate development of new approaches to the tasks. Many early efforts focused on collecting large-scale annotations to support data-driven approaches for low-level tasks such as POS tagging (Marcus, Santorini, & Marcinkiewicz, 1993) and named entity recognition (Grishman & Sundheim, 1996). Other early benchmarks, such as the Message Understanding Conference (MUC) benchmarks for information extraction (Sundheim, 1993) and the automatic content extraction (ACE) benchmarks for coreference resolution and relation extraction (Doddington, Mitchell, Przybocki, Ramshaw, Strassel, & Weischedel, 2004), were mainly designed to explore the information found in linguistic knowledge and context. In part motivated by earlier works from McCarthy (1976) and Lehnert (1977), other benchmarks were created to focus on reading comprehension through tasks like question answering (Hirschman, Light, Breck, & Burger, 1999; Voorhees, 2000). These early reading comprehension benchmarks often required only one sentence of context at a time to answer questions(Kočiský, Schwarz, Blunsom, Dyer, Hermann, Melis, & Grefenstette, 2018), possible through shallow linguistic approaches. Although it may be the case that some type of outside knowledge or reasoning may be required to reach an oracle performance, as all of these earlier benchmarks primarily targeted approaches that only applied linguistic context, they are outside the scope of this survey.

Recent years have seen a shift in benchmark tasks which are beyond the use of linguistic context, aiming to require a deeper understanding to solve the tasks. For instance, consider this question from the Winograd Schema Challenge (Levesque, 2011):"The trophy would not fit in the brown suitcase because it was too big. What was too big?" To answer this question, linguistic constraints will not be able to resolve whether *it* refers to *the trophy* or *the brown suitcase*. Only based on external knowledge (i.e., the fact that an object must be bigger than another object in order to contain it) is it possible to resolve the pronoun *it* and answer the question correctly. Higher-level tasks, such as question answering, may appear the same as in earlier benchmarks. However, successfully solving these tasks will also require machines to go beyond linguistic context, and rely on reasoning and knowledge that is not explicitly stated in text. In this survey, we attempt to provide an overview of recent benchmarks like these.

These benchmarks vary in terms of the scope of tasks, for example, from the specific and focused reference resolution task in the Winograd Schema Challenge (Levesque, 2011) to broader tasks such as textual entailment, e.g., the RTE Challenges (Dagan, Glickman, & Magnini, 2005). While some benchmarks only focus on one specific task, others are comprised of a variety of tasks, e.g., GLUE (Wang, Singh, Michael, Hill, Levy, & Bowman, 2018). Some benchmarks target a specific type of knowledge, e.g., commonsense social psychology in Event2Mind (Rashkin, Sap, Allaway, Smith, & Choi, 2018b), while others intend to address a variety of types of knowledge, e.g., the Story Cloze Test (Mostafazadeh, Chambers, He, Parikh, Batra, Vanderwende, Kohli, & Allen, 2016),

which requires a wide range of facts about everyday life. How benchmark tasks are formulated also differs among existing benchmarks. Some are in the form of multiple-choice questions, requiring a binary decision or a selection from a candidate list, while others are more open-ended. Different characteristics of benchmarks serve different goals, and a critical question is what criteria to consider in order to create a benchmark that can support technology development and measurable outcome of research progress on deep reasoning abilities.

Section 2 gives a detailed account of existing benchmarks and their common characteristics. It also summarizes important criteria to consider in building these benchmarks, including factors to consider in task formulation, techniques to avoid problematic data bias, data collection methods, and factors affecting overall benchmark complexity.

1.2 Scope of Knowledge Resources

Humans perform natural language inference based on a vast amount of external knowledge about language and the world. To support machines' inference ability, compiled knowledge resources become important. It is an unsolved problem to comprehensively taxonomize all of the human knowledge and reasoning required to perform general inference (Davis & Marcus, 2015). For the purpose of this survey, though, we attempt to provide an overview of three types of knowledge resources: *linguistic knowledge*, *common knowledge*, and *commonsense knowledge*. Although modern approaches often rely on large text corpora to implicitly incorporate external knowledge without regard to what types of knowledge are required, we think it would be helpful to distinguish different types of knowledge to facilitate better understanding of the nature of NLI tasks.

Linguistic knowledge. In order to comprehend human language, machines first need *linguistic knowledge*, i.e., knowledge about the language. This includes an understanding of word meanings, grammar, syntax, semantics, and discourse structure. Having linguistic knowledge gives a human or machine the basic capabilities of understanding language, and is a required property of virtually any NLP system, even those not created for NLI tasks.

Common knowledge. *Common knowledge* refers to well-known facts about the world that are often explicitly stated, e.g., "canine distemper is a domestic animal disease" (Cambria, Song, Wang, & Hussain, 2011). This kind of knowledge is often referred to in human communication (Cambria et al., 2011). Some types of common knowledge may be domain-specific, e.g., the knowledge required to answer questions on an elementary-level science test (Sugawara, Yokono, & Aizawa, 2017). While domain-specific knowledge is obviously useful for domain-specific applications, much of this knowledge may not be needed for general-purpose communication with humans.

Commonsense knowledge. *Commonsense knowledge*, on the other hand, is typically unstated, as it is considered obvious to most humans (Cambria et al., 2011), and consists of universally accepted beliefs about the world (Nunberg, 1987). Davis and Marcus (2015) demonstrate this: "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, you do not have to ask which is which." Commonsense knowledge provides a deeper understanding of language. While it is rarely referred to in language, humans rely on it in communication (Cambria et al., 2011), as it is required to reach a common ground (Chai, Gao, She, Yang, Saba-Sadiya, & Xu, 2018). Commonsense knowledge consists of everyday assumptions about the world (Zang, Cao, Cao, Wu, & CAO, 2013), and is generally learned through one's own experience with the world, but can also be inferred by generalizing over common knowledge (Speer,

Havasi, & Lieberman, 2008). While common knowledge can vary by region, culture, and other factors, we expect that commonsense knowledge should be roughly typical to all humans (Davis, 2017).

Two particularly important domains of commonsense knowledge are intuitive physics, i.e., humans' basic understanding of physical interactions in the world, and intuitive psychology, i.e., humans' basic understanding of human emotion, behavior, and motives (DARPA, 2018). Also respectively referred to as naïve physics and psychology (Davis & Marcus, 2015) or physical and social commonsense (Sakaguchi, Bras, Bhagavatula, & Choi, 2019), these are considered fundamental as they have been shown to develop in humans as infants, even before language develops (Baillargeon, 2004; Song, Baillargeon, & Fisher, 2005), and humans are capable of several kinds of intuitive physical and psychological inference at only eighteen months of age (DARPA, 2018). These types of knowledge allow us to make on-the-fly predictions about everyday physical and social interactions. Intuitive physics allows us to infer, for example, the potential consequences when dropping a glass of water on the floor, i.e., the glass may shatter, water will almost certainly spill on the floor, and the floor could become slippery. From intuitive psychology, we can further infer that dropping the glass was likely an accident, and the person who dropped the glass will be angry, upset, or embarrassed.

Section 3 gives an introduction to existing knowledge resources, and several recent efforts in building such resources to facilitate natural language inference.

1.3 Scope of Approaches

To tackle the challenging benchmark tasks, many computational models have been developed. These range from earlier symbolic and statistical approaches to recent approaches based on deep neural networks. Especially for recent neural approaches, various architectures have been proposed that model context of language, take advantage of external data or knowledge resources, and achieve the state-of-the-art performance, and in some cases, achieve near or above human performance. However, due to the unexplainability of these recent approaches, as well as statistical biases recently found in benchmark datasets (Schwartz, Sap, Konstas, Zilles, Choi, & Smith, 2017; Gururangan, Swayamdipta, Levy, Schwartz, Bowman, & Smith, 2018; Niven & Kao, 2019), it remains a subject of debate in terms of how much progress we have made in enabling natural language inference ability. Section 4 summarizes recent learning and inference approaches, and discusses their performance, limitations, and trends.

2. Benchmarks and Tasks

Alan Turing perhaps is the first one who proposed an experiment, widely-known as the Turing Test (Turing, 1950), to evaluate whether a machine possessed intelligence. The formulation of the Turing Test, however, has been criticized for encouraging machines to deceive humans (Levesque, 2011), not providing feedback on a continuous scale to allow for incremental development (Ortiz, 2016), and being impractical in several ways (Harnad, 2000). Benchmark datasets have thus been used as alternatives to the Turing Test, providing training and testing data, evaluation frameworks, and continuous numerical feedback for various language tasks. For the remainder of this section, we first give an overview of recent benchmarks that are designed to evaluate machine's inference ability in natural language understanding. Then, we summarize important considerations in creating these benchmarks and lessons learned from ongoing research.
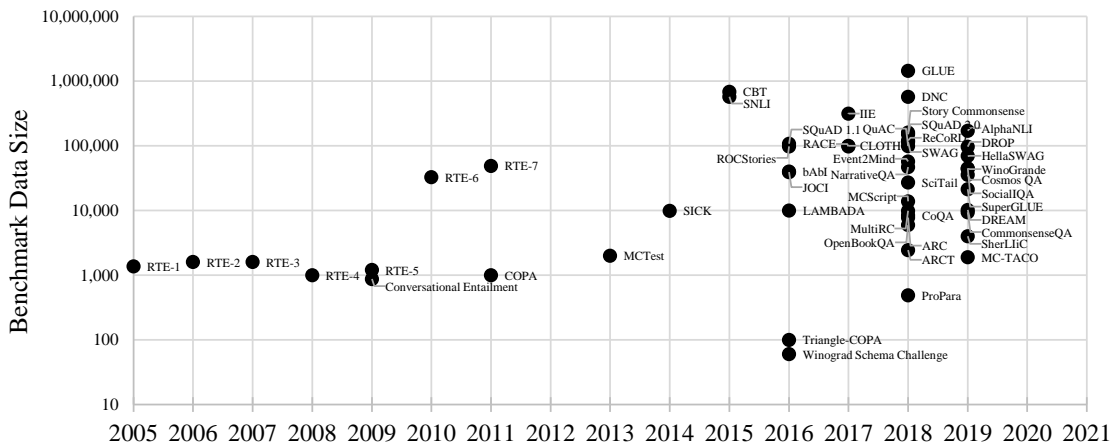
Figure 2: Since the early 2000s, there has been a surge of benchmark tasks geared toward natural language inference. In 2018, we saw the creation of more benchmarks of larger sizes than ever before.

2.1 An Overview of Existing Benchmarks

Since the Recognizing Textual Entailment (RTE) Challenges introduced by Dagan et al. (2005) in the early 2000s, there has been an explosion of challenging NLI benchmarks being created. Figure 2 shows a trend of this growth among the benchmarks introduced in this section. The RTE challenges, which were inspired by difficulties encountered in semantic processing for machine translation, question answering, and information extraction (Dagan et al., 2005), had been the dominant reasoning tasks for many years. Recently, however, there has been a surge in the variety of benchmarks. As shown in Figure 2, the sizes for most benchmarks in earlier days, e.g., from 2005 to 2015, are relatively small. Most benchmarks developed after 2015 have a much larger number of instances. Some of them have more than 100,000 instances, which make the application of deep learning approaches possible.

Next, we give a review of widely used benchmarks, introduced by the following groupings: reference resolution, question answering, textual entailment, plausible inference, psychological reasoning, and multiple tasks. These groupings are not necessarily exclusive, but they highlight the recent trends in the field. While all benchmarks' training and/or development data are available for free download, it is important to note that test data are often not distributed publicly so that testing of systems can occur privately in a standard, unbiased way.

2.1.1 REFERENCE RESOLUTION

Reference resolution is the process of identifying a referent, typically a linguistic mention in a span of text, that a particular expression, e.g., a pronoun or phrase, refers to. This process can be significantly complicated by ambiguities which arise from the presence of multiple entities in a sentence with pronouns, creating a need for external knowledge, e.g., commonsense knowledge, to inform decisions (Levesque, 2011). Since such challenging examples for reference resolution are typically handcrafted or handpicked (Morgenstern, Davis, & Ortiz, 2016; Rahman & Ng, 2012; Rudinger, Naradowsky, Leonard, & Van Durme, 2018a), there exists a small magnitude of data for

this skill, and it remains an unsolved problem (Rahman & Ng, 2012; Davis, Morgenstern, & Ortiz, 2017). Consequently, there is a need for more benchmarks here. In the following paragraphs, we introduce the few benchmarks that currently exist for reference resolution.

Winograd Schema Challenge.   The classic reference resolution benchmark is the Winograd Schema Challenge (WSC), inspired by Winograd (1972), originally proposed by Levesque (2011), later developed by Levesque, Davis, and Morgenstern (2012), and actualized by Morgenstern and Ortiz (2015) and Morgenstern et al. (2016). In this challenge, systems are presented with questions about sentences known as Winograd schemas. To answer a question, a system must disambiguate a pronoun whose referent may be one of two entities, and can be changed by replacing a single word in the sentence. Consequently, linguistic context is typically not useful in disambiguating the pronoun, and outside knowledge is required. The first Challenge dataset consisted of just 60 testing examples (Morgenstern, 2016), but there are more available online.[4] Additionally, Rahman and Ng (2012) curated nearly 1,000 similar pronoun resolution problems to aid in training systems for the WSC; this set of problems is included within a later-mentioned dataset, and thus is not introduced separately. Further information about the Winograd Schema Challenge is available at `https://commonsensereasoning.org/winograd.html`.

WinoGrande.   Inspired by biases found in the original Winograd Schema Challenge (Trichelair, Emami, Trischler, Suleman, & Cheung, 2018), Sakaguchi et al. (2019) created WinoGrande, a large-scale, adversarial version of the problem focusing primarily on intuitive physics and psychology. It consists of nearly 44,000 examples, but maintains the same level of difficulty, having low baseline performance but high human performance. This is the first high-quality, large-scale dataset of adversarial Winograd schemas, and should encourage the application of deep neural networks to the problem to improve machine performance on reference resolution tasks. WinoGrande can be downloaded at `https://mosaic.allenai.org/projects/winogrande`.

Other reference resolution benchmarks.   Other reference resolution benchmarks include the basic and compound coreference subtasks within bAbI (Weston, Bordes, Chopra, Rush, van Merriënboer, Joulin, & Mikolov, 2016), the definite pronoun resolution problems by Rahman and Ng (2012) within Inference is Everything (White, Rastogi, Duh, & Van Durme, 2017), the Winograd NLI subtask within GLUE (Wang et al., 2018), and the gendered anaphora-based Winogender task originally by Rudinger et al. (2018a) within DNC (Poliak, Haldar, Rudinger, Hu, Pavlick, White, & Van Durme, 2018a). The full benchmarks to which these subtasks belong are introduced in Section 2.1.6, while in Figure 3, we list several examples of challenging reference resolution problems across the existing benchmarks, where external knowledge, such as the knowledge that predators eat their prey, is useful in disambiguating pronouns.

It is important to note that all of these benchmarks evaluate the resolution of entity references rather than event references. Benchmarks toward event coreference are limited, and those which do exist typically require only linguistic context to disambiguate event references (Doddington et al., 2004; Bejan & Harabagiu, 2008). On the other hand, since the linguistic structures which refer to specific events are much more flexible than those for entities (Chen & Ji, 2009), disambiguating event references may be more challenging than the existing benchmarks for entity reference resolu-

---

4. See `https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`,
   `https://commonsensereasoning.org/winograd.html`, and
   `https://commonsensereasoning.org/disambiguation.html`.

(A) **Winograd Schema Challenge (Levesque, 2011)**

The trophy would not fit in the brown suitcase because it was too big. What was too big?

    a. **The trophy**

    b. The suitcase

The trophy would not fit in the brown suitcase because it was too small. What was too small?

    a. The trophy

    b. **The suitcase**

(B) **Winogender (Rudinger et al., 2018a)**

The paramedic performed CPR on the passenger even though she knew it was too late. Who knew it was too late?

    a. **The paramedic**

    b. The passenger

(C) **(Rahman & Ng, 2012)**

Lions eat zebras because they are predators. Who are predators?

    a. **Lions**

    b. Zebras

Figure 3: Examples from existing reference resolution benchmark tasks. Answers in bold.

tion. As more progress is made, we anticipate event reference resolution benchmarks which require outside knowledge and reasoning will be created in the near future.

2.1.2 QUESTION ANSWERING

Instead of providing a focused language processing task like reference resolution, many benchmarks instead provide a more comprehensive mix of language processing and reasoning skills within a single task. Question answering (QA) is one such comprehensive task, particularly the recent formulation of the task providing a passage, and requiring a system to answer questions about it to demonstrate its comprehension of the passage (also referred to as machine reading comprehension). Today, QA is a fairly well-established area in NLP, and there are many existing benchmarks for QA. However, not all of them require external knowledge and reasoning to solve (Ostermann, Modi, Roth, Thater, & Pinkal, 2018), so many can be solved only based on linguistic context without a deep understanding. Some QA benchmarks that do require external knowledge are MCScript (Ostermann et al., 2018), CoQA (Reddy, Chen, & Manning, 2018), and OpenBookQA (Mihaylov, Clark, Khot, & Sabharwal, 2018). Examples from these benchmarks are listed in Figure 4, where the commonsense knowledge that diapers are typically thrown away, and the common knowledge that steel is a conductor and that Democrats and Republicans compete in American elections, is particularly useful in answering questions. Other QA benchmarks require advanced reasoning processes beyond the text. SQuAD 2.0 is one such example, as it includes unanswerable questions about passages (Rajpurkar, Jia, & Liang, 2018), which require a deep reasoning over all knowledge given in the passage to identify. In the following paragraphs, we introduce all surveyed QA benchmarks.

MCTest. MCTest by Richardson, Burges, and Renshaw (2013) is a collection of 500 fictional stories and 2,000 multiple-choice questions which require outside knowledge and various types of

(A) MCScript (Ostermann et al., 2018)
Did they throw away the old diaper?

    a. **Yes, they put it into the bin.**

    b. No, they kept it for a while.

(B) OpenBookQA (Mihaylov et al., 2018)
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.

Evidence: Metal is a thermal conductor.

(C) CoQA (Reddy et al., 2018)
The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Who is the democratic candidate?
Terry McAuliffe
Evidence: Democrat Terry McAuliffe

Who is his opponent?
Ken Cuccinelli
Evidence: Republican Ken Cuccinelli

Figure 4: Examples from QA benchmarks which require inference through outside knowledge. Answers in bold.

reasoning to answer. More than half of the questions require more than one sentence to answer, so that systems must synthesize information from multiple locations rather than using shallow matching techniques to find the answer. The questions are open-domain with limited restriction on what will be asked, but stories and questions were designed so that they can be understood by young children, thus limiting the amount of required outside knowledge to be somewhat manageable. MCTest may be downloaded at `https://github.com/mcobzarenco/mctest/tree/master/data/MCTest`.

RACE. The Reading Comprehension Dataset from Examinations (RACE) by Lai, Xie, Liu, Yang, and Hovy (2017) consists of about 98,000 multiple-choice reading comprehension questions over 28,000 passages collected from middle- and high-school English exams. As the exams are designed by expert teachers, the questions are challenging, often requiring reasoning over multiple sentences in the passage. Further, many questions and candidate choices cannot be matched directly to the passage. RACE can be requested for download at `https://www.cs.cmu.edu/~glai1/data/race/`.

NarrativeQA. NarrativeQA by Kočiský et al. (2018) is a collection of about 1,500 full stories and movie scripts paired with human-written summaries, along with about 47,000 human-written questions and answers based on the summaries. Two open-ended tasks are proposed: answering the questions paired with the original stories and scripts, and answering the questions paired with the summaries. As the questions were written based upon summaries of the stories and scripts, answering the questions based on the full stories and scripts is more challenging, and requires reasoning over a larger span of text. NarrativeQA can be downloaded at `https://github.com/deepmind/narrativeqa`.

ARC. The AI2 Reasoning Challenge (ARC) from Clark, Cowhey, Etzioni, Khot, Sabharwal, Schoenick, and Tafjord (2018) provides a dataset of almost 8,000 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. As many of the questions require systems to draw information from multiple sentences in the corpus to answer correctly, it is not possible to accurately solve the task by simply searching the corpus for keywords. As such, this task encourages reasoning over a large amount of knowledge, an essential skill for NLI systems. ARC can be downloaded at `https://data.allenai.org/arc/`.

MCScript. The MCScript benchmark by Ostermann et al. (2018) is one of few QA benchmarks which emphasizes commonsense knowledge. The dataset consists of about 14,000 2-way multiple-choice questions based on short passages, with a large proportion of its questions requiring pure commonsense knowledge to answer, and thus can be answered without the passage. Such an example is given in Figure 4. Questions are conveniently labeled with whether they are answered with the provided text or commonsense. A download link to MCScript can be found at `https://www.sfb1102.uni-saarland.de/?page_id=2582`.

ProPara. ProPara by Mishra, Huang, Tandon, Yih, and Clark (2018) consists of 488 annotated paragraphs of procedural text. These paragraphs describe various processes such as photosynthesis and hydroelectric power generation in order for systems to learn object tracking in processes which involve changes of state. The authors assert that recognizing these state changes can require knowledge of physical processes, so proficiency in reasoning with such knowledge is required to perform

well. Annotations of the paragraphs are in the form of a grid which describes the state of each participant in the paragraph after each sentence of the paragraph. If a system understands a paragraph in this dataset, it is said that for each entity mentioned in the paragraph, it can answer any question about whether the entity is created, destroyed, or moved, and when and where this happens. To answer all possible questions accurately, a system must produce a grid identical to the annotations for the paragraph. Thus, systems are evaluated by their performance on this task. ProPara data is linked from `https://data.allenai.org/propara/`.

MultiRC. MultiRC by Khashabi, Chaturvedi, Roth, Upadhyay, and Roth (2018) is a reading comprehension dataset consisting of about 10,000 questions posed on over 800 paragraphs across a variety of topic domains. It differs from a traditional reading comprehension dataset in that most questions can only be answered by reasoning over multiple sentences in the accompanying paragraphs, 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 entirely variable. All of these features make it difficult for shallow approaches to perform well on the benchmark, and encourage a deeper understanding of the passage. Further, the benchmark includes a variety of non-trivial semantic phenomena in passages, such as coreference and causal relationships, which often require external knowledge to recognize and reason about. MultiRC can be downloaded at `https://cogcomp.org/multirc/`.

ARCT. The Argument Reasoning Comprehension Task (ARCT) proposed by Habernal, Wachsmuth, Gurevych, and Stein (2018) is a different type of question answering task which provides the structures of arguments presented in online comments on news articles, requiring systems to choose one of two lexically similar warrants to justify a claim made in an argument. According to the authors, many such warrants, e.g., "most house cats face enemies," are unstated, rather presupposed by arguments, and require external knowledge to infer. This formulation makes available some of the required knowledge for interpreting the presented arguments, however many of the examples require further world knowledge beyond the provided warrants (Niven & Kao, 2019). ARCT consists of about 2,500 examples, and is available at `https://github.com/UKPLab/argument-reasoning-comprehension-task`.

SQuAD. The Stanford Question Answering Dataset (SQuAD) from Rajpurkar, Zhang, Lopyrev, and Liang (2016) was originally a set of about 100,000 questions posed on passages from Wikipedia articles, which are provided with the questions. The initial dataset did not require any external knowledge or reasoning to solve; answers were spans of text directly from the passage, and could be found based on linguistic context. To make this dataset more challenging, SQuAD 2.0 (Rajpurkar et al., 2018) was later released to add about 50,000 additional questions which are unanswerable given the passage. Such questions require one to reason about what can and cannot be inferred given the knowledge from the passage, which takes a deep understanding. All SQuAD data can be downloaded at `https://rajpurkar.github.io/SQuAD-explorer/`.

CoQA. The Conversational Question Answering (CoQA) dataset by Reddy et al. (2018) contains passages each accompanied with a set of questions in conversational form, as well as their answers and evidence for the answers as spans of text. There are about 127,000 questions in the dataset total, where answers are abstractive but a rationale must be given as a span of text. As questions are conversational, each question pertaining to a passage must be answered given the conversation history generated from previous questions on the passage. Unlike other conver-

sational QA datasets, CoQA explicitly requires external knowledge in many questions, e.g., the example given in Figure 4. Further, CoQA includes out-of-domain types of questions appearing only in the test set, and unanswerable questions throughout, encouraging reasoning beyond the passage and knowledge learned in training to answer questions. CoQA data can be downloaded at `https://stanfordnlp.github.io/coqa/`.

QuAC. Question Answering in Context (QuAC) by Choi, He, Iyyer, Yatskar, Yih, Choi, Liang, and Zettlemoyer (2018) is a similar but larger conversational QA benchmark consisting of about 98,000 questions over nearly 14,000 dialogs about nearly 9,000 passages. Answers here are spans of text, while in CoQA answers are abstractive, but rationalized by spans of text. Unlike CoQA, when generating conversations, the worker asking the questions could not see the text and thus did not know the answer, encouraging the creation of questions which cannot be answered by string matching approaches. To ensure productive conversations, the worker answering the questions provided dialog acts which gave the other worker feedback on whether to follow up the previous question or change topics. Like CoQA, QuAC includes unanswerable questions, but does not explicitly address the use of external knowledge. QuAC can be downloaded at `https://quac.ai/`.

OpenBookQA. OpenBookQA by Mihaylov et al. (2018) intends to address some shortcomings of previous QA datasets. Earlier datasets often do not require external knowledge or reasoning to solve, and those that do require vast domains of knowledge which are difficult to capture. OpenBookQA contains about 6,000 4-way multiple choice science questions which may require science facts (common knowledge) or commonsense knowledge, e.g., the example in Figure 4. Instead of providing no knowledge resource like MCScript (Ostermann et al., 2018), or a prohibitively large corpus of facts to support answering the questions like ARC (Clark et al., 2018), OpenBookQA provides an "open book" of about 1,300 science facts to support answering the questions, each associated directly with the question(s) they apply to. For required commonsense knowledge, the authors expect that outside resources will be used in answering the questions. Instructions to download OpenBookQA can be found at `https://github.com/allenai/OpenBookQA`.

CommonsenseQA. CommonsenseQA by Talmor, Herzig, Lourie, and Berant (2019) is another QA benchmark which directly targets commonsense knowledge, consisting of 9,500 three-way multiple-choice questions. To ensure commonsense knowledge is used to answer questions, each question requires one to disambiguate a target concept from three connected concepts in ConceptNet, a popular commonsense knowledge graph (Liu & Singh, 2004). Utilizing a large, general-purpose knowledge graph like ConceptNet ensures not only that questions target commonsense relations directly, but that the domain of commonsense knowledge required by questions is fairly comprehensive for everyday use. CommonsenseQA data can be downloaded at `https://www.tau-nlp.org/commonsenseqa`.

DREAM. The Dialogue-Based Reading Comprehension Examination (DREAM) by Sun, Yu, Chen, Yu, Choi, and Cardie (2019) also targets the use of external knowledge. The benchmark consists of about 10,000 multiple-choice questions posed on about 6,500 selections of dialogue between multiple parties. Their analysis shows that the answers to most questions cannot be directly extracted from the text, and will require multi-sentence reasoning. Further, 34% of questions explicitly require external knowledge to answer. Such questions are challenging to current neural reading comprehension systems, especially those which are not typically used for processing dialogue. DREAM can be downloaded at `https://dataset.org/dream/`.

DROP. The Discrete Reasoning Over Paragraphs (DROP) benchmark by Dua, Wang, Dasigi, Stanovsky, Singh, and Gardner (2019) consists of nearly 100,000 questions over about 7,000 passages from Wikipedia. Unlike previous similar benchmarks, however, the questions in DROP deliberately require systems to perform various reasoning tasks, including arithmetic (e.g., addition, subtraction), comparison, and more. This promotes a deep understanding of passages in order to use the knowledge in them for reasoning tasks. Further, the data was adversarially created, only including questions which a baseline neural system could not answer correctly. DROP can be downloaded at `https://allennlp.org/drop`.

Cosmos QA. Commonsense Machine Comprehension (Cosmos) QA by (Huang, Bras, Bhagavatula, & Choi, 2019) consists of 35,600 multiple-choice reading comprehension questions, each based on an accompanying text. Questions are designed to require external knowledge and reasoning, some questions are unanswerable, and most correct answers are not stated in the context passage. About 94% of questions require commonsense knowledge, which is the highest proportion seen yet for any QA benchmark requiring the comprehension of a passage. Cosmos QA can be downloaded at `https://wilburone.github.io/cosmos/`.

MC-TACO. The Multiple Choice Temporal Common-sense (MC-TACO) benchmark by Zhou, Khashabi, Ning, and Roth (2019) provides about 1,900 multiple-choice questions requiring commonsense knowledge about the temporal properties of events, e.g., their duration, frequency, and order. Such commonsense knowledge was under-emphasized in previous benchmarks. Questions can have multiple correct answers. MC-TACO can be downloaded at `https://leaderboard.allenai.org/mctaco/submissions/public`.

2.1.3 TEXTUAL ENTAILMENT

Textual entailment is defined by Dagan et al. (2005) as a directional relationship between a text and a hypothesis, where it can be said that the text *entails* the hypothesis if a typical person would infer that the hypothesis is true given the text. Some benchmarks expand this task by also requiring recognition of contradiction, e.g., the fourth and fifth RTE Challenges (Giampiccolo, Dang, Magnini, Dagan, & Dolan, 2008; Bentivogli, Dagan, Dang, Giampiccolo, & Magnini, 2009). Like question answering, this task requires the utilization of several simpler language processing skills, such as named entity recognition and coreference resolution. Unlike question answering, since it also requires a sense of what a typical person would infer, commonsense knowledge is inherently essential to the textual entailment task. The RTE Challenges (Dagan et al., 2005) are the classic benchmarks for entailment, but there are now several larger benchmarks inspired by them. Examples from these benchmarks are listed in Figure 5, and they require common knowledge such as the concept of a snow angel, and commonsense knowledge such as the relationship between the presence of a crowd and loneliness. In the following paragraphs, we introduce several textual entailment benchmarks in detail.

RTE Challenges. The Recognizing Textual Entailment (RTE) Challenge introduced the task of textual entailment, and aimed to evaluate a machines' acquisition of the common background knowledge and reasoning capabilities a typical human needs to determine whether one text entails another (Dagan et al., 2005). The inaugural Challenge provided a task where given a text and

---

5. Example extracted from the SICK data available at `https://clic.cimec.unitn.it/composes/sick.html`.

(A) RTE Challenge (Dagan et al., 2005)
Text: American Airlines began laying off hundreds of flight attendants on Tuesday, after a federal judge turned aside a union's bid to block the job losses.
Hypothesis: American Airlines will recall hundreds of flight attendants as it steps up the number of flights it operates.

Label: **not entailment**

(B) SICK (Marelli, Menini, Baroni, Bentivogli, Bernardi, & Zamparelli, 2014a)[5]
Sentence 1: Two children are lying in the snow and are drawing angels.
Sentence 2: Two children are lying in the snow and are making snow angels.

Label: **entailment**

(C) SNLI (Bowman et al., 2015)
Text: A black race car starts up in front of a crowd of people.
Hypothesis: A man is driving down a lonely road.

Label: **contradiction**

(D) MultiNLI, Telephone (Williams, Nangia, & Bowman, 2017)
Context: that doesn't seem fair does it
Hypothesis: There's no doubt that it's fair.

Label: **contradiction**

(E) SciTail (Khot, Sabharwal, & Clark, 2018)
Premise: During periods of drought, trees died and prairie plants took over previously forested regions.
Hypothesis: Because trees add water vapor to air, cutting down forests leads to longer periods of drought.

Label: **neutral**

Figure 5: Examples from RTE benchmarks. Answers in bold.

hypothesis, systems were expected to predict whether the text entailed the hypothesis. In following years, more similar Challenges took place (Bar-Haim, Dagan, Dolan, Ferro, Giampiccolo, Magnini, & Szpektor, 2006; Giampiccolo, Magnini, Dagan, & Dolan, 2007). The fourth and fifth Challenge added a new three-way decision task where systems were additionally expected to recognize contradiction relationships between texts and hypotheses (Giampiccolo et al., 2008; Bentivogli et al., 2009). The main task for the sixth and seventh Challenges instead provided one hypothesis and several potential entailing sentences in a corpus (Bentivogli, Clark, Dagan, & Giampiccolo, 2010, 2011). The eighth Challenge (Dzikovska, Nielsen, Brew, Leacock, Giampiccolo, Bentivogli, Clark, Dagan, & Dang, 2013) addressed a bit different problem which focused on classifying student responses as an effort toward providing automatic feedback in an educational setting. The first five RTE Challenge datasets consisted of around 1,000 examples each (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009), while the sixth and seventh consisted of about 33,000 and 49,000 examples respectively. Data from all RTE Challenges can be downloaded at `https://tac.nist.gov/`.

Conversational Entailment. The Conversational Entailment dataset, introduced by Zhang and Chai (2009) and further analyzed by Zhang and Chai (2010), is a small collection of 875 binary entailment pairs consisting of a premise from a natural language dialogue and a hypothesis that requires interpreting the dialogue. Many examples require predicting dialogue participants' beliefs, desires, and intentions, which need to be reasoned from the context of dialogue. The dataset can be downloaded at `https://github.com/lairmsu/ConversationalEntailment`.

SICK. The Sentences Involving Compositional Knowledge (SICK) benchmark by Marelli et al. (2014a) is a collection of almost 10,000 pairs of sentences. Two tasks are presented for the pairs: sentence relatedness and entailment. The entailment task, which is more related to our survey, is a 3-way decision task in the style of RTE-4 (Giampiccolo et al., 2008) and RTE-5 (Bentivogli et al., 2009). SICK can be downloaded at `https://clic.cimec.unitn.it/composes/sick.html`.

SNLI. The Stanford Natural Language Inference (SNLI) benchmark from Bowman et al. (2015) contains nearly 600,000 sentence pairs, and also provides a 3-way decision task similar to the fourth and fifth RTE Challenges (Giampiccolo et al., 2008; Bentivogli et al., 2009). In addition to gold labels for entailment, contradiction, or neutral, the SNLI data includes five crowd judgments for the label, which may indicate a level of confidence or agreement for it. This benchmark is later expanded into MultiNLI (Williams et al., 2017), which follows the same format, but includes sentences of various genres, such as telephone conversations. MultiNLI is included within the later introduced GLUE benchmark (Wang et al., 2018), while SNLI can be downloaded at `https://nlp.stanford.edu/projects/snli/`.

SciTail. SciTail by Khot et al. (2018) consists of about 27,000 premise-hypothesis sentence pairs adapted from science questions into a 2-way entailment task. Unlike other entailment benchmarks, this one is primarily science-based, which may require domain-specific knowledge. SciTail can be downloaded at `https://data.allenai.org/scitail/`.

SherLIiC. SherLIiC by Schmitt and Schütze (2019) is a benchmark for lexical inference in context (LIiC), i.e., a subset of textual entailment focusing on relationships between words, which Glockner, Shwartz, and Goldberg (2018) found that previous benchmarks lacked, and state-of-the-art systems struggled with. SherLIiC focuses particularly on entailment relationships between

event and action verbs. It consists of about 960,000 inference rule candidates formed from pairs of relations extracted from a knowledge graph, each typed for better word disambiguation. Almost 4,000 of these are annotated for their correctness. Additionally, the benchmark includes several automatic calculations of similarity between premises and hypotheses. SherLIiC serves as a challenging testbed for NLI and textual entailment systems. It can be downloaded at `https://github.com/mnschmit/SherLIiC`.

2.1.4 PLAUSIBLE INFERENCE

While textual entailment benchmarks require one to draw concrete conclusions, others require hypothetical, intermediate, or uncertain conclusions based upon a limited context, defined as plausible inference by Davis and Marcus (2015), and historically referred to as abductive reasoning (Peirce, 1883; Hobbs, Stickel, Appelt, & Martin, 1993). Performing this process requires reasoning over linguistic context and external knowledge. Popular task formulations include cloze tasks, i.e., fill-in-the-blank tasks, and sentence and story completion. Examples from these benchmarks are listed in Figure 6, and they require commonsense knowledge of everyday interactions, e.g., answering a door or proposing for marriage, and activities, e.g., making scrambled eggs or being snowed in. In the following paragraphs, we introduce all surveyed plausible inference benchmarks.

COPA. The Choice of Plausible Alternatives (COPA) task by Roemmele et al. (2011) evaluates causal reasoning between events, which requires commonsense knowledge about what usually takes place in the world. Each example provides a premise and either asks for the correct cause or effect from two choices, thus testing either backward or forward causal reasoning. COPA data, which consists of 1,000 examples total, can be downloaded at `https://people.ict.usc.edu/~gordon/copa.html`.

CBT. The Children's Book Test (CBT) from Hill, Bordes, Chopra, and Weston (2015) consists of about 687,000 cloze-style questions about 20-sentence stories mined from publicly available children's books. Each question requires a system to fill a blank in a new line following the story given a set of 10 candidate words. Questions are classified into 4 tasks based on the type of the missing word to predict, which can be a named entity, common noun, verb, or preposition. To predict a named entity or common noun, an understanding of the story and the actions involving the entities in it will be required. To predict a verb or preposition, outside linguistic and common knowledge may be useful. CBT can be downloaded at `https://research.fb.com/downloads/babi/`.

ROCStories. ROCStories by Mostafazadeh et al. (2016) is a corpus of about 50,000 five-sentence everyday life stories, containing a host of causal and temporal relationships between events, ideal for learning commonsense knowledge. Of the stories, about 3,700 are designated as test cases, which include a plausible and implausible alternate story ending for trained systems to choose between. The task of solving the ROCStories test cases is called the Story Cloze Test, a more challenging alternative to the narrative cloze task proposed by Chambers and Jurafsky (2008). The most recent release of ROCStories, which can be requested at `https://cs.rochester.edu/nlp/rocstories/`, adds about 50,000 stories to the dataset.

LAMBADA. The Language Modeling Broadened to Account for Discourse Aspects (LAMBADA) benchmark by Paperno, Kruszewski, Lazaridou, Pham, Bernardi, Pezzelle, Baroni, Boleda, and Fernandez (2016) is an open-ended cloze task which consists of about 10,000 passages from BooksCorpus (Zhu, Kiros, Zemel, Salakhutdinov, Urtasun, Torralba, & Fidler, 2015) where a missing tar-

(A) COPA (Roemmele, Bejan, & Gordon, 2011)

I knocked on my neighbor's door. What happened as result?

    a. **My neighbor invited me in.**

    b. My neighbor left his house.

(B) JOCI (Zhang, Rudinger, Duh, & Van Durme, 2017)

Context: John was excited to go to the fair
Hypothesis: The fair opens.
Label: **5 (very likely)**

Context: Today my water heater broke
Hypothesis: A person looks for a heater.
Label: **4 (likely)**

Context: John's goal was to learn how to draw well
Hypothesis: A person accomplishes the goal.
Label: **3 (plausible)**

Context: Kelly was playing a soccer match for her University
Hypothesis: The University is dismantled.
Label: **2 (technically possible)**

Context: A brown-haired lady dressed all in blue denim sits in a group of pigeons.
Hypothesis: People are made of the denim.
Label: **1 (impossible)**

(C) ROCStories (Mostafazadeh et al., 2016)

Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.

Ending:

    a. **Tom asked Sheryl to marry him.**

    b. He wiped mud off of his boot.

(D) AlphaNLI (Bhagavatula, Bras, Malaviya, Sakaguchi, Holtzman, Rashkin, Downey, Yih, & Choi, 2019)

Observation 1: There was ten feet of snow outside.
Observation 2: In all that time I was unable to check my mail.
Hypotheses:

    a. **I couldn't open my door against a drift for 3 days.**

    b. It took 10 minutes for the snow plow to come through.

(E) SWAG (Zellers, Bisk, Schwartz, & Choi, 2018)

He pours the raw egg batter into the pan. He

    a. drops the tiny pan onto a plate

    b. **lifts the pan and moves it around to shuffle the eggs.**

    c. stirs the dough into a kite.

    d. swirls the stir under the adhesive.

Figure 6: Examples from benchmarks requiring plausible inference. Answers in bold.

get word is predicted in the last sentence of each passage. This is similar to CBT Hill et al. (2015), but here, the missing word is constrained to always be the last word of the last sentence, there are no candidate words to choose from, and most importantly, examples were filtered by humans to ensure they were possible to guess given the context, i.e., the sentences in the passage leading up to the last sentence. Examples were further filtered to ensure that missing words could not be guessed without the context, ensuring that models attempting the dataset would need to reason over the entire paragraph to answer questions. Further, in a later study on a random sampling of 100 passages from LAMBADA, 24 were found to require external knowledge Chu, Wang, Gimpel, and McAllester (2017). LAMBADA can be downloaded at `https://zenodo.org/record/2630551#.XS5JHuhKhPZ`.

JOCI. The JHU (John Hopkins University) Ordinal Commonsense Inference (JOCI) benchmark by Zhang et al. (2017) consists of about 39,000 sentence pairs, each consisting of a context and hypothesis. Given these, systems must rate how likely the hypothesis is on a scale from 1 to 5, where 1 corresponds to impossible, 2 to technically possible, 3 to plausible, 4 to likely, and 5 to very likely. This task is similar to SNLI (Bowman et al., 2015) and other 3-way entailment tasks, but provides more options which essentially range between entailment and contradiction. This is fitting, considering the fuzzier nature of the plausible inference task compared to textual entailment. JOCI can be downloaded from `https://github.com/sheng-z/JOCI`.

CLOTH. The Cloze Test by Teachers (CLOTH) benchmark by Xie, Lai, Dai, and Hovy (2017) is a collection of nearly 100,000 4-way multiple-choice cloze-style questions from middle- and high school-level English language exams, where the answer fills a blank in a given text. Each question is labeled with a type of deep reasoning it involves, where the four possible types are grammar, short-term reasoning, matching/paraphrasing, and long-term reasoning, i.e., reasoning over multiple sentences. CLOTH data can be downloaded at `https://www.cs.cmu.edu/~glai1/data/cloth/`.

SWAG. Situations with Adversarial Generations (SWAG) from Zellers et al. (2018) is a benchmark dataset of about 113,000 beginnings of small texts each with four possible endings. Given the context each text provides, systems decide which of the four endings is most plausible in a task referred to as commonsense NLI. To ensure the benchmark was challenging to state-of-the-art systems and could not be solved through shallow approaches or by exploiting data bias, examples were filtered with a new process called adversarial filtering (explained further in Section 6). Examples include labels for the source of the correct ending, and ordinal labels for the likelihood of each possible ending and the correct ending. SWAG data can be downloaded at `https://rowanzellers.com/swag/`.

ReCoRD. The Reading Comprehension with Commonsense Reasoning (ReCoRD) benchmark by Zhang, Liu, Liu, Gao, Duh, and Van Durme (2018), similar to SQuAD (Rajpurkar et al., 2016), consists of questions posed on passages, particularly news articles. However, questions in ReCoRD are in cloze format, requiring more hypothetical reasoning, and many questions explicitly require commonsense knowledge and reasoning over multiple sentences to answer. Named entities are identified in the data, and are used to fill the blanks for the cloze task. The benchmark data consist of over 120,000 examples, most of which are claimed to require commonsense reasoning. ReCoRD can be downloaded at `https://sheng-z.github.io/ReCoRD-explorer/`.

HellaSWAG. HellaSWAG by Zellers, Holtzman, Bisk, Farhadi, and Choi (2019a) is a reissue of SWAG (Zellers et al., 2018) consisting of 70,000 examples in the same format, but created from different sources, i.e., ActivityNet captions (Heilbron, Escorcia, Ghanem, & Niebles, 2015) along with WikiHow texts rather than text from the Large Scale Movie Description Challenge (Rohrbach, Torabi, Rohrbach, Tandon, Pal, Larochelle, Courville, & Schiele, 2017). Further, negative endings are generated with a better language model, and again adversarially filtered to be challenging for state-of-the-art models that were created since SWAG's first release. HellaSWAG can be downloaded at `https://rowanzellers.com/hellaswag/`.

AlphaNLI. AlphaNLI by Bhagavatula et al. (2019) is an abductive reasoning task where given two observations as an incomplete context, one must predict which of two hypothesized events is more plausible to have happened between the observations. This is similar to the Story Cloze Test (Mostafazadeh et al., 2016), which AlphaNLI is partially constructed from. However, rather than inferring the most plausible ending of an otherwise complete story, one must infer the most plausible explanation of a situation given two partial observations of it. As such, the benchmark closely resembles the classical formulation of abductive reasoning studied with formal logic (Peirce, 1883; Hobbs et al., 1993), unlike other benchmarks in this category, and performing the task requires reasoning with unstated commonsense implications. AlphaNLI consists of about 20,000 possible contexts and 200,000 hypotheses, paired to form about 170,000 examples in the training and development sets.[6] AlphaNLI can be downloaded at `https://leaderboard.allenai.org/anli/submissions/get-started`.

2.1.5 INTUITIVE PSYCHOLOGY

An especially significant domain in plausible inference tasks is intuitive psychology, as inference of emotions and intentions through behavior is a fundamental capability of humans (Gordon, 2016). Several benchmarks touch on this topic in some examples, e.g., the marriage proposal example in ROCStories (Mostafazadeh et al., 2016) from Figure 6, but some are entirely focused here. Examples from each benchmark are listed in Figure 7, and they require intuitive social and psychological commonsense knowledge such as plausible reactions to being punched or yelled at. In the following paragraphs, we introduce these benchmarks in detail.

Triangle-COPA. Triangle-COPA by Gordon (2016) is a variation of COPA (Roemmele et al., 2011) based on a past social psychology experiment by Heider and Simmel (1944). It contains 100 examples in the format of COPA, and accompanying videos. Questions focus specifically on emotions, intentions, and other aspects of social psychology. The data also includes logical forms of the questions and alternatives, as the paper focuses on logical formalisms for psychological commonsense knowledge and reasoning. Triangle-COPA can be downloaded at `https://github.com/asgordon/TriangleCOPA`.

Story Commonsense. As mentioned earlier, the stories from ROCStories by Mostafazadeh et al. (2016) contain sociological and psychological instances of commonsense knowledge. Motivated by classical theories of motivation and emotions from psychology, Rashkin et al. (2018a) cre-

---

6. The authors do not report the magnitude of any of the dataset partitions, so these counts came from the publicly available partitions of the dataset.

7. Example extracted from Story Commonsense development data available at `https://uwnlp.github.io/storycommonsense/`.

(A) **Triangle-COPA (Gordon, 2016)**

A circle and a triangle are in the house and are arguing. The circle punches the triangle. The triangle runs out of the house. Why does the triangle leave the house?

    a. The triangle leaves the house because it wants the circle to come fight it outside.

    b. **The triangle leaves the house because it is afraid of being further assaulted by the circle.**

(B) **SC (Rashkin, Bosselut, Sap, Knight, & Choi, 2018a)**[7]

Jervis has been single for a long time.
He wants to have a girlfriend.
One day he meets a nice girl at the grocery store.

Motivation: **to be loved, companionship**
Emotions: **shocked, excited, hope, shy, fine, wanted**

Maslow: **love**
Reiss: **contact, romance**
Plutchik: **joy, surprise, anticipation, trust, fear**

(C) **Event2Mind (Rashkin et al., 2018b)**

PersonX starts to yell at PersonY

PersonX's intent: **to express anger, to vent their frustration, to get PersonY's full attention**
PersonX's reaction: **mad, frustrated, annoyed**
Other's reactions: **shocked, humiliated, mad at PersonX**

(D) **SocialIQA (Sap, Rashkin, Chen, Le-Bras, & Choi, 2019b)**

Tracy accidentally pressed upon Austin in the small elevator and it was awkward. Why did Tracy do this?

    a. get very close to Austin

    b. **squeeze into the elevator**

    c. get flirty with Austin

Figure 7: Examples from intuitive psychology benchmarks. Answers in bold.

ated the Story Commonsense benchmark containing about 160,000 annotations of the motivations and emotions of characters in ROCStories to enable more concrete reasoning in this area. In addition to the tasks of generating motivational and emotional annotations, the dataset introduces three classification tasks: one for inferring the basic human needs theorized by Maslow (1943), one for inferring the human motives theorized by Reiss (2004), and one for inferring the human emotions theorized by Plutchik (1980). Story Commonsense can be downloaded at `https://uwnlp.github.io/storycommonsense/`.

Event2Mind. In addition to motivations and emotions, systems may need to infer intentions and reactions surrounding events. To support this, Rashkin et al. (2018b) introduce Event2Mind, a benchmark dataset of about 57,000 annotations of intentions and reactions for about 25,000 unique events extracted from other corpora, including ROCStories (Mostafazadeh et al., 2016). Each event involves one or two participants, and presents three tasks of predicting the primary participant's (1) intentions and (2) reactions, and predicting (3) the reactions of other participants. Event2Mind can be downloaded at `https://uwnlp.github.io/event2mind/`.

SocialIQA. Social Intelligence QA (SocialIQA) by Sap et al. (2019b) is a multiple-choice question answering benchmark containing 45,000 questions requiring intuitive psychology and commonsense knowledge of social interactions. Each question consists of a brief context, a question about the context, and three answer options. Interestingly, the authors show that when fine-tuned on SocialIQA, deep learning models perform better when later fine-tuned on smaller-magnitude benchmarks which require commonsense knowledge such as COPA (Roemmele et al., 2011) and the Winograd Schema Challenge (Davis et al., 2017), exceeding the state of the art on both. This suggests that some useful commonsense knowledge can be acquired from the data. More information about SocialIQA can be found at `https://maartensap.github.io/social-iqa/`.

Task	bAbI	IIE	GLUE	DNC	SuperGLUE
Semantic Role Labeling		✓			
Relation Extraction	✓			✓	
Event Factuality				✓	
Named Entity Recognition				✓	
Word Sense Disambiguation					✓
Reference Resolution	✓	✓	✓	✓	✓
Grammaticality			✓		
Lexicosyntactic Inference				✓	✓
Sentiment Analysis			✓	✓	
Figurative Language				✓	
Sentence Similarity			✓		
Paraphrase		✓	✓	✓	
Sentence Completion					✓
Textual Entailment			✓	✓	✓
Question Answering	✓		✓		✓

Table 1: Comparison of language processing tasks present in the bAbI (Weston et al., 2016), IIE (White et al., 2017), GLUE (Wang et al., 2018), DNC (Poliak et al., 2018a), and SuperGLUE (Wang et al., 2019) benchmarks. Recent multi-task benchmarks focus on the recognition of a greater variety of linguistic and semantic phenomena.

2.1.6 MULTIPLE TASKS

Some benchmarks consist of several focused language processing or reasoning subtasks so that a diverse set of reading comprehension skills can be learned and tested in a consistent format. These benchmarks can be used as diagnostics to determine how a model performs in different areas (Wang et al., 2018). Subtasks are often reframed from various pre-existing benchmarks (White et al., 2017). As shown in Table 1, many benchmark subtasks are specified by low- to high-level language processing tasks. In the following paragraphs, we introduce these benchmarks in detail.

bAbI. The bAbI benchmark from Weston et al. (2016) 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. Each task also focuses on a different type of reasoning or language processing task, allowing systems to learn basic skills one at a time. Tasks are as follows:

1. Single supporting fact
2. Two supporting facts
3. Three supporting facts
4. Two argument relations
5. Three argument relations
6. Yes/no questions
7. Counting
8. Lists/sets
9. Simple negation
10. Indefinite knowledge
11. Basic coreference
12. Conjunction
13. Compound coreference
14. Time reasoning
15. Basic deduction
16. Basic induction
17. Positional reasoning
18. Size reasoning
19. Path finding
20. Agent's motivations

In addition to providing focused language processing tasks as previously discussed, bAbI also provides focused commonsense reasoning tasks, such as its tasks for deduction and induction, and time, positional, and size reasoning (Weston et al., 2016). It is important to note that bAbI's data is synthetic and very simply structured, and the inference required for testing data can be learned mostly from the training data, so it is easily solved by current systems. Further, models trained on bAbI currently cannot generalize well to real-world, naturally-generated data (Das, Munkhdalai, Yuan, Trischler, & McCallum, 2019). As such, bAbI is best used to gauge a model's ability to learn reasoning skills rather than its comprehension of natural language. A dataset by Nematzadeh, Burns, Grant, Gopnik, and Griffiths (2018) expands bAbI with questions about theory of mind, which may make it a more realistic reasoning task. bAbI can be downloaded at `https://research.fb.com/downloads/babi/`.

Inference is Everything. Inference is Everything (IIE) by White et al. (2017) follows bAbI (Weston et al., 2016) in creating a suite of tasks, where each task is deliberately geared toward a different language processing task: semantic proto-role labeling, paraphrase, and pronoun resolution. Each task is in classic RTE Challenge format (Dagan et al., 2005), i.e., given context and hypothesis texts, one must determine whether the context entails the hypothesis. Between the tasks, IIE includes about 300,000 examples, all of which are recast from previously existing datasets. IIE can be downloaded with another multi-task suite at `https://github.com/decompositional-semantics-initiative/DNC`.

GLUE. The General Language Understanding Evaluation (GLUE) dataset from Wang et al. (2018) consists of 9 language tasks, including single-sentence binary classification and 2- or 3-way entailment comparable to the dual tasks in RTE-4 and RTE-5 (Giampiccolo et al., 2008; Bentivogli et al., 2009). The GLUE tasks are recast or included directly from other benchmark data and corpora:

- Corpus of Linguistic Acceptability (CoLA) from Warstadt, Singh, and Bowman (2018)
- Stanford Sentiment Treebank (SST-2) from Socher, Perelygin, Wu, Chuang, Manning, Ng, and Potts (2013)
- Microsoft Research Paraphrase Corpus (MRPC) from Dolan and Brockett (2005)
- Quora Question Pairs (QQP) from Iyer, Dandekar, and Csernai (2017)
- Semantic Textual Similarity Benchmark (STS-B) from Cer, Diab, Agirre, Lopez-Gazpio, and Specia (2017)
- Multi-Genre Natural Language Inference (MNLI) from Williams et al. (2017)
- Question Natural Language Inference (QNLI) recast from SQuAD 1.1 Rajpurkar et al. (2016)
- Recognizing Textual Entailment (RTE), consisting of examples from RTE-1 (Dagan et al., 2005), RTE-2 (Bar-Haim et al., 2006), RTE-3 (Giampiccolo et al., 2007), and RTE-5 (Bentivogli et al., 2009)

- Winograd Natural Language Inference (WNLI) recast from privately shared Winograd schemas by the creators of the Winograd Schema Challenge Levesque (2011)

Many of these tasks, though not all, require some outside knowledge and reasoning to make inferences. GLUE includes a small analysis set for diagnostic purposes, which has manual annotations of fine-grained knowledge and reasoning categories examples fall into. Overall, GLUE has over 1 million examples, which can be downloaded at `https://gluebenchmark.com/tasks`.

DNC. The Diverse Natural Language Inference Collection (DNC) by Poliak et al. (2018a) consists of 9 textual entailment tasks requiring 7 different types of reasoning. Like IIE by White et al. (2017), data for each task follows the form of the original RTE Challenge (Dagan et al., 2005). Some of the tasks within DNC cover basic language processing tasks like named entity recognition, while others cover more challenging tasks like anaphora resolution and recognition of puns, which require some outside knowledge and reasoning beyond the text. Each task is recast from a previously existing dataset:

1. Event Factuality, recast from UW (Lee, Artzi, Choi, & Zettlemoyer, 2015), MEANTIME (Minard, Speranza, Urizar, Altuna, van Erp, Schoen, & van Son, 2016), and (Rudinger, White, & Van Durme, 2018b)
2. Named Entity Recognition, recast from the Groningen Meaning Bank (Bos, Basile, Evang, Venhuizen, & Bjerva, 2017) and the ConLL-2003 shared task (Tjong Kim Sang & De Meulder, 2003)
3. Gendered Anaphora Resolution, recast from the Winogender dataset (Rudinger et al., 2018a)
4. Lexicosyntactic Inference, recast from MegaVeridicality (White & Rawlins, 2018), VerbNet (Schuler, 2005), and VerbCorner (Hartshorne, Bonial, & Palmer, 2013)
5. Figurative Language, recast from puns by Yang, Lavie, Dyer, and Hovy (2015) and Miller, Hempelmann, and Gurevych (2017)
6. Relation Extraction, partially from FACC1 (Gabrilovich, Ringgaard, & Subramanya, 2013)
7. Subjectivity, recast from Kotzias, Denil, De Freitas, and Smyth (2015)

The DNC benchmark consists of about 570,000 examples total, and can be downloaded at `https://github.com/decompositional-semantics-initiative/DNC`.

SuperGLUE. SuperGLUE by Wang et al. (2019) is a reissue of GLUE (Wang et al., 2018) with about 23,000 examples over a different selection of more difficult reframed single-sentence binary classification and 2- or 3-way entailment tasks:

- CommitmentBank (CB) from de Marneffe, Simons, and Tonhauser (2019)
- Choice of Plausible Alternatives (COPA) from Roemmele et al. (2011)
- Multi-Sentence Reading Comprehension (MultiRC) from Khashabi et al. (2018)
- Recognizing Textual Entailment (RTE), consisting of examples from RTE-1 (Dagan et al., 2005), RTE-2 (Bar-Haim et al., 2006), RTE-3 (Giampiccolo et al., 2007), and RTE-5 (Bentivogli et al., 2009)
- Word-in-Context (WiC) from Pilehvar and Camacho-Collados (2019)
- Winograd Schema Challenge (WSC) from Levesque (2011) recast instead into a single-sentence binary classification task[8]

---

8. Consists of original Winograd Schema Challenge examples (Morgenstern et al., 2016), examples by the authors available online, and examples privately shared by the authors.

Many of these tasks require reasoning over multiple sentences, and using external knowledge to perform language processing tasks like word-sense disambiguation and reference resolution. SuperGLUE can be downloaded at `https://super.gluebenchmark.com/`.

2.2 Criteria and Considerations for Creating Benchmarks

The goal of benchmarks is to support technology development and provide a platform to measure research progress. Whether this goal can be achieved depends on the nature of the benchmark. This section identifies the successes and lessons learned from the creation of the surveyed benchmarks, and summarizes key considerations and criteria that should guide the creation of the benchmarks, particularly in the areas of task format, evaluation schemes, data biases, data collection methods, and benchmark complexity.

2.2.1 TASK FORMAT

In creating benchmarks, determining the formulation of the problem is an important step. Among existing benchmarks, there exist a few common task formats, and while some formats are interchangeable, others are only suited for particular tasks. We provide a review of these formats, indicating the types of tasks they are suitable for.

Classification tasks. Most benchmark tasks are classification problems, where each response is a single choice from a finite number of options. These include textual entailment tasks, which most commonly require a binary or ternary decision about a pair of sentences, cloze tasks, which require a multiple-choice decision to fill in a blank, and traditional multiple-choice question answering tasks.

Textual entailment tasks. A highly popular format was originally introduced by the RTE Challenges, where given a pair of texts, i.e., a context and hypothesis, one must determine whether the context entails the hypothesis (Dagan et al., 2005). In the fourth and fifth RTE Challenges, this format was extended to a three-way decision problem where the hypothesis may contradict the context (Giampiccolo et al., 2008; Bentivogli et al., 2009). The JOCI benchmark further extends the problem to a five-way decision task, where the hypothesis text may range from impossible to very likely given the context (Zhang et al., 2017).

While this format is typically used for textual entailment problems like the RTE Challenges, it can be used for nearly any type of inference problem. Some multi-task benchmarks have adopted the format for several different reasoning tasks, for example, the Inference is Everything (White et al., 2017), GLUE (Wang et al., 2018), and DNC (Poliak et al., 2018a) benchmarks use the classic RTE format for most of their subtasks by automatically recasting previously existing benchmarks into it. Many of these problems deal either with reasoning processes more focused than the RTE Challenges, or more advanced reasoning than the RTE Challenges demand, showing the flexibility of the format. Some such tasks include reference resolution, recognition of puns, and question answering. Examples from these recasted subtasks are listed in Figure 8.

Cloze tasks. Another popular format is the cloze task, originally conceived by Taylor (1953). Such a task typically involves the deletion of one or more words in a text, essentially requiring one to fill in the blank, usually from a set of choices, but not always, e.g., LAMBADA, which is an open-ended cloze task (Paperno et al., 2016). This format expands upon the traditional language modeling task, i.e., the task of predicting the next word in a sequence of words, inspired by the need to calculate a probability distribution of sequences in language for problems like speech recognition

---

(A) GLUE, Question Answering NLI

Context: Who was the main performer at this year's halftime show?

Hypothesis: The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively.

Label: **entailed**

(B) IIE, Definite Pronoun Resolution

Context: The bird ate the pie and it died.

Hypothesis: The bird ate the pie and the bird died.

Label: **entailed**

(C) DNC, Figurative Language

Context: Carter heard that a gardener who moved back to his home town rediscovered his roots.

Hypothesis: Carter heard a pun.

Label: **entailed**

Figure 8: Examples from tasks within the General Language Understanding Evaluation (GLUE) benchmark by Wang et al. (2018), Inference is Everything (IIE) by White et al. (2017), and the Diverse NLI Collection (DNC) by Poliak et al. (2018a). Each task is recast from preexisting data into classic RTE format.

(Bahl, Jelinek, & Mercer, 1983) and machine translation (Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Lafferty, Mercer, & Roossin, 1990). The cloze format has been used in several recent NLI benchmarks, including CBT (Hill et al., 2015), the Story Cloze Test for the ROCStories benchmark (Mostafazadeh et al., 2016), CLOTH (Xie et al., 2017), SWAG (Zellers et al., 2018), and ReCoRD (Zhang et al., 2018). These recent benchmarks provide anywhere from two to ten options to fill in the blank, and range from requiring the prediction of a single word to parts of sentences and entire sentences. Examples of cloze tasks are listed in Figure 9.

Traditional multiple-choice tasks. If not in entailment or cloze form, classification tasks tend to be formulated as traditional multiple-choice questions. Benchmarks which use this format include COPA (Roemmele et al., 2011), the Winograd Schema Challenge (Davis, Morgenstern, & Ortiz, 2018), and MCScript (Ostermann et al., 2018). Two-way and four-way decision questions are the most common among the surveyed multiple-choice benchmarks.

Open-ended tasks. On the other hand, some benchmarks require more open-ended responses rather than providing a small list of alternatives to choose from. Answers may be restricted to spans of a given text, e.g., SQuAD (Rajpurkar et al., 2016, 2018) or QuAC (Choi et al., 2018). They may be less restricted to a subset of a large number of category labels, e.g., the Maslow, Reiss, and Plutchik tasks in Story Commonsense (Rashkin et al., 2018a). And of course, they may be purely open-ended, e.g., Event2Mind (Rashkin et al., 2018b), NarrativeQA (Kočiský et al., 2018), or DROP (Dua et al., 2019). Examples of these open-ended formats are listed in Figure 10.

---

9. Example extracted from Story Commonsense test data available at `https://uwnlp.github.io/storycommonsense/`

(A) CBT (Hill et al., 2015)

1 Mr. Cropper was opposed to our hiring you .

2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .

3 He says female teachers ca n't keep order .

4 He 's started in with a spite at you on general principles , and the boys know it .

5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .

6 Cropper is sly and slippery , and it is hard to corner him . "

7 " Are the boys big ? "

8 queried Esther anxiously .

9 " Yes .

10 Thirteen and fourteen and big for their age .

11 You ca n't whip 'em – that is the trouble .

12 A man might , but they 'd twist you around their fingers .

13 You 'll have your hands full , I 'm afraid .

14 But maybe they 'll behave all right after all . "

15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best .

16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .

17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .

18 He was a big , handsome man with a very suave , polite manner .

19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .

20 Esther felt relieved .

21 She thought that Mr. _____ had exaggerated matters a little .

Blank: **Baxter**, Cropper, Esther, course, fingers, manner, objection, opinions, right, spite

(B) ROCStories (Mostafazadeh et al., 2016)

Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

Ending:

a. **Karen became good friends with her roommate.**

b. Karen hated her roommate.

(C) CLOTH (Xie et al., 2017)

She pushed the door open and found nobody there. "I am the _____ to arrive." She thought and came to her desk.

a. last

b. second

c. third

d. **first**

(D) SWAG (Zellers et al., 2018)

On stage, a woman takes a seat at the piano. She

a. sits on a bench as her sister plays with the doll.

b. smiles with someone as the music plays.

c. is in the crowd, watching the dancers.

d. **nervously sets her fingers on the keys.**

(E) ReCoRD (Zhang et al., 2018)

... Daniela Hantuchova knocks Venus Williams out of Eastbourne 6-2 5-7 6-2 ...

Query: Hantuchova breezed through the first set in just under 40 minutes after breaking Williams' serve twice to take it 6-2 and led the second 4-2 before _____ hit her stride.

Venus Williams

Figure 9: Examples from cloze tasks. Answers in bold.

(A) SQuAD 2.0 (Rajpurkar et al., 2018)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

(B) bAbI (Weston et al., 2016)

The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.

How do you go from den to kitchen?
west, north

(C) SC (Rashkin et al., 2018a)[9]

Valerie was getting ready for a formal dance. She had been preparing for hours. As she was ready to leave, her acrylic nail broke. She snapped off all of her faux nails.

Maslow: **esteem, stability**
Reiss: **status, approval, order**
Plutchik: **surprise, sadness, disgust, anger**

(D) DROP (Dua et al., 2019)

That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate.

Question: How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?
4300000

Figure 10: Examples from open-ended response tasks. Answers in bold.

2.2.2 EVALUATION SCHEMES

As mentioned earlier, the Turing Test (Turing, 1950) has long been criticized by AI researchers as it does not seem to truly evaluate machine intelligence. As described by Ortiz (2016), there is a critical need for new intelligence benchmarks to support incremental development and evaluation of AI techniques. These benchmarks should not merely provide a pass or fail grade, rather they should provide feedback on a continuous scale which enables both incremental development and comparison of approaches. One key consideration for these benchmarks is informative evaluation metrics that are objective and easy to calculate. These metrics can be used to compare different approaches and compare machine performance against human performance.

Evaluation metrics. Choice of evaluation metrics is highly dependent on the type of task, and thus so is the difficulty of calculating them. Classification tasks often use exact-match accuracy if correct answers or class labels are evenly distributed through benchmark data. If this is not the case, common practice is to additionally present F-measure as an evaluation metric (Wang et al., 2018). The precision and recall may also be presented, however the F-measure (which considers both in its calculation) is much more common in the recent surveyed benchmarks. Multiple-choice and classification task formats such as RTE, cloze, and traditional multiple-choice can all use these metrics.

More open-ended tasks are by nature more difficult to evaluate, but they can still be objective and informative. Tasks like SQuAD (Rajpurkar et al., 2016) or QuAC (Choi et al., 2018), where answers can only be spans of a provided text, can be evaluated similarly to multiple-choice tasks. Exact-match accuracy and F-measure are used as evaluation metrics on both of these benchmarks, where the collection of tokens in the predicted and true spans (excluding punctuation and articles) are compared. Where answers are instead a subset of a large group of category labels, e.g., the Maslow, Reiss, and Plutchik tasks in Story Commonsense, evaluation is similar, but for these benchmarks particularly, precision and recall are additionally calculated (Rashkin et al., 2018a).

Where multiple purely open-ended responses are given, as in Event2Mind (Rashkin et al., 2018b), evaluation is more difficult. Event2Mind particularly uses the average cross-entropy and the "recall @ 10," i.e., the percentage of times human-produced ground truth labels fall within the top 10 generated predictions from a system. In bAbI (Weston et al., 2016), LAMBADA (Paperno et al., 2016), and CoQA (Reddy et al., 2018), where open-ended responses are compared to a single ground truth answer, exact-match accuracy and F-measure are used. These benchmarks are able to use such exact evaluation measures because responses are short. In bAbI particularly, correct responses are limited to one word or lists of words, and in LAMBADA, responses are only a single word to fill a blank. Such restrictions are essential for such simple and accurate evaluation of open-ended responses. For longer generated responses, e.g., as required in NarrativeQA (Kočiský et al., 2018), evaluation metrics for machine translation like BLEU (Papineni, Roukos, Ward, & Zhu, 2002) or Meteor (Denkowski & Lavie, 2011), or similar metrics for text summarization like ROUGE (Lin, 2004), are often used to compare answers with gold responses. However, many such metrics have been found to not correlate well with human judgments of generated text (Belz & Reiter, 2006; Elliott & Keller, 2014), so they should be interpreted with caution.

Baseline performance. Benchmarks provide common datasets and experimental setups for researchers to compare different approaches. When a new benchmark is first released, results from simple baseline approaches are often reported. Ideally, the benchmark should be designed in a

way that a simple baseline would achieve poor performance to ensure that it cannot be solved by trivial means, instead requiring a more advanced approach. For classification problems, simple baseline approaches are often calculated by random choice, choosing the majority class from the training data, or choosing the alternative with the highest overlap in n-grams with the question or provided text (Richardson et al., 2013). Examples of these baseline approaches can be found in the Story Cloze Test baselines, which include most of these approaches and more (Mostafazadeh et al., 2016). For open-ended problems, shallow lexical approaches (e.g., using language models) may again be used, as in the bAbI benchmark(Weston et al., 2016). When releasing a new benchmark, results from one or more state-of-the-art models for similar existing benchmarks will also typically be reported as baselines. Later, as more advanced approaches are developed for the benchmark, they are often used as state-of-the-art baselines on the next generation of benchmarks. This cycle continually boosts baseline performance over time, and drives the development of new and improved models as well as new benchmarks.

Human performance measurement. To evaluate the progress of machine intelligence, human performance on benchmark tasks is often measured to provide a reference point. Human performance may be measured in a few different ways. A typical way is to select several human judges, which may be crowdsourced, to solve a random subset of the benchmark data. Performance may be averaged across judges, e.g., as done for CBT (Hill et al., 2015), or calculated through a majority vote of all judges, e.g., for SWAG (Zellers et al., 2018). Human performance can also be measured through annotator agreement, e.g., as done for ROCStories (Mostafazadeh et al., 2016). An "expert" human performance measurement may be additionally provided as a ceiling, typically by testing researchers involved with creating the dataset. This was done, for example, by the authors of RACE (Lai et al., 2017) and SWAG.

In many benchmarks, one measure for computational models is to assess how close it comes to or whether it exceeds human performance. When a system exceeds human performance, this seems to suggest that the system has acquired the skills and knowledge required by the benchmark. However, due to the unexplainability of many neural models, this is often difficult to prove. Further, data biases have recently been found in many popular benchmarks (Gururangan et al., 2018), allowing both simple approaches (Schwartz et al., 2017) and state-of-the-art neural systems (Niven & Kao, 2019) to exploit the biases and achieve artificially high performance. This puts these unexplainable systems under higher scrutiny.

2.2.3 DATA BIASES

When creating benchmarks, one challenge is the bias of data unintentionally introduced to the benchmark. For example, in the first release of the Visual Question Answering (VQA) benchmark (Agrawal, Lu, Antol, Mitchell, Zitnick, Parikh, & Batra, 2017), researchers found that machine learning models were learning several statistical biases in the data, and could answer up to 48% of questions in the validation set without seeing the image (Manjunatha, Saini, & Davis, 2018). This artificially high system performance is problematic, as it is not accredit to the underlying technology. Here, we summarize several key dimensions of biases encountered in previous NLI research. Some of these (e.g., class label distributions) are easier to avoid, while others (e.g., hidden correlation biases) are more difficult to address.

Label distribution bias. Class label distribution bias is the easiest to avoid. In multiple-choice problems, ordinal positions of correct answers should be entirely randomized so that each possible

choice appears in benchmark data in a uniform distribution. For other classification tasks where class labels have specific meanings, e.g., textual entailment, ensuring balance of class labels must occur earlier, i.e., when creating examples. Care should be taken to ensure an equal number of examples are generated for each class. When class labels are balanced, a majority-class baseline will score as low as possible on the task. While binary-choice tasks should have a 50% majority class baseline, the MegaVeridicality subtask within DNC (Poliak et al., 2018a) has a 67% majority-class baseline due to unevenly distributed class labels, leaving significantly less room for incremental improvement than benchmarks with lower-performing simple baselines.

Question type bias. For benchmarks involving question answering, previous work has made effort to balance the types of questions, especially if questions are generated by crowdsourcing. This will ensure a broad domain of knowledge and reasoning required to solve the benchmark. A fairly simple method to keep a balance of question types is to calculate the distribution of the first words of each question, as the creators of CoQA (Reddy et al., 2018) and CommonsenseQA (Talmor et al., 2019) did. One could also manually label and analyze a random sample of questions with categories relating to the types of knowledge or reasoning required, or have crowd workers perform the labeling if an expert is not essential. Examples of this are shown by the creators of SQuAD 2.0 (Rajpurkar et al., 2018) and ARC (Clark et al., 2018). To entirely avoid question type biases, implementing a standard set of questions for all provided texts may be beneficial. ProPara does this for all entities in its procedural paragraphs (Mishra et al., 2018), limiting questions about each entity to whether it is created, destroyed, or moved during the paragraph, and when and where this happens. Manjunatha et al. (2018) suggest that biases can further be avoided in VQA benchmarks by forcing questions to require a particular skill (e.g., telling time or subtracting) to be answered. This rule of thumb can be applicable for textual benchmarks as well, e.g., in the DROP benchmark (Dua et al., 2019), which requires discrete reasoning skills like arithmetic to answer questions about passages.

Superficial correlation bias. The kind of biases most difficult to discern and avoid perhaps are those caused by accidental correlations between features of answers and questions. One example of this is gender bias, which NLI systems are particularly vulnerable to when trained on biased data. Rudinger et al. (2018a) highlight this problem in coreference resolution, showing that systems trained on gender-biased data perform worse in gender pronoun disambiguation. For example, consider the problem from their Winogender dataset in Figure 3: "The paramedic performed CPR on the passenger even though she knew it was too late." In determining who *she* is, systems trained on gender-biased training data may be more likely, for example, to incorrectly choose *the passenger* rather than *the paramedic* due to male gender pronouns appearing in training data more commonly in the context of this occupation than female gender pronouns. To avoid this, gender pronouns should appear equally frequently among other words in training data, especially those related to occupations and activities. Similar gender biases are identified in Event2Mind data, which are derived from movie scripts (Rashkin et al., 2018b).

When authoring natural language data (e.g., generating questions or hypotheses), some human stylistic artifacts such as predictable sentence structure, presence of certain linguistic phenomena, and vocabulary use can also cause these superficial correlation biases. This is most often the case if data is authored by crowd workers (Gururangan et al., 2018), but similar biases have been found in expert-authored data as well (Trichelair et al., 2018). In the Story Cloze Test (Mostafazadeh et al., 2016), systems are presented a plausible and implausible ending to the story, and must choose which ending is plausible. However, previous work Schwartz et al. (2017) has shown that the Story Cloze

Test can be solved with up to 72.4% accuracy by only looking at the two possible endings. They do this by exploiting human writing style biases in the possible endings rather than performing a more realistic reasoning. For example, they find that negative language is used more commonly in the wrong ending (e.g., "hates"), and the correct ending is more likely to use enthusiastic language (e.g., "!"). An example of a biased negative ending is seen in Figure 9. Sharma, Allen, Bakhshandeh, and Mostafazadeh (2018) have begun work to update the benchmark data and remove these biases.

While generating the Story Cloze Test data was not a fast or simple task for crowd workers, Gururangan et al. (2018) suggest that such biases can come from crowd workers' adoption of predictable annotation strategies and heuristics to quickly generate data. These strategies have been revealed for several textual entailment benchmarks which consist of pairs of short sentences. For example, on the entailment task in SemEval 2014 (Marelli, Bentivogli, Baroni, Bernardi, Menini, & Zamparelli, 2014b) as part of the SICK benchmark (Marelli et al., 2014a), Lai and Hockenmaier (2014) found that the presence of negation in an example was associated with the appearance of the contradiction class label. Their trained classifier using this feature alone was able to achieve 61% accuracy. Later, Poliak, Naradowsky, Haldar, Rudinger, and Van Durme (2018b) and Gururangan et al. (2018) found the presence of particular words in the hypothesis sentence can bias the entailment prediction in several entailment benchmarks. For example, "nobody" in contradictory examples from SNLI (Bowman et al., 2015) was found to be an indicator of contradiction, while generic words like "animal" and "instrument", as well as gender-neutral pronouns, were found to be indicators of entailment. Gururangan et al. further find that a high sentence length is an indicator of a neutral entailment relationship, and suggest that crowd workers often remove words from the context sentences to create entailed hypothesis sentences. Using biases like these, a baseline approach by Poliak et al. which only used the hypothesis sentence from entailment benchmarks was able to outperform a majority-class baseline in SNLI, JOCI (Zhang et al., 2017), SciTail (Khot et al., 2018), two tasks within Inference is Everything (White et al., 2017), and the MultiNLI task within GLUE (Williams et al., 2017; Wang et al., 2018).

Addressing superficial correlation bias. Various approaches have been developed to address this data bias problem. For example, to recognize biases, a simple technique is to calculate the mutual information between words and prediction classes within benchmark examples. This was performed by researchers in discovering stylistic biases in entailment benchmarks (Gururangan et al., 2018). This type of analysis should be performed on any new benchmark data when it is created.

To avoid the biases, more involved techniques may be required. For example, in creating the SWAG benchmark (Zellers et al., 2018), a novel adversarial filtering process was introduced to ensure writing styles are consistent among ending choices, and the correct answer cannot be identified by exploitative stylistic classifiers. This is done by over-generating negative choices for each question with a language model, then using a collection of strong classifiers to filter out negative examples which are too easy to identify given an arbitrary train-test split. When SWAG was released, it was challenging to state-of-the-art classifiers. But it did not take long before a new model solved SWAG with a high accuracy (Devlin, Chang, Lee, & Toutanova, 2018). By incorporating this new model into the adversarial filtering, as well as using a newer language model to generate text, Zellers et al. (2019a) created an updated version called HellaSWAG, on which the new state-of-the-art model struggled. Variants of adversarial filtering were used in creating other benchmarks such as WinoGrande (Sakaguchi et al., 2019) and AlphaNLI (Bhagavatula et al., 2019) to respectively remove statistical biases and select the most challenging pairs of choices from a pool of hypotheses,

resulting in low performance from state-of-the-art baseline approaches. As this iterative process of discovering stronger classifiers and in turn strengthening the adversarial filtering continues, an important question will be when all of the statistical biases are removed, or if it will ever be possible to remove all statistical biases. Nonetheless, given these recent successes, adversarial filtering should continue to be used on new benchmarks to challenge the latest state-of-the-art models, and perhaps be used to strengthen existing benchmarks where possible.

Another technique to avoid these biases proposed by Belinkov, Poliak, Shieber, Van Durme, and Rush (2019), particularly in textual entailment problems, is to train a model to predict the premise sentence given the hypothesis sentence and class label. This training setup ensures that the model ignores unexpected language artifacts in the premise, attempting to remove superficial statistical biases. While the development of these paradigms to avoid data bias is promising, a continuous effort in finding and improving techniques to avoid these kinds of biases will be important as the state of the art continues to evolve.

2.2.4 COLLECTION METHODS

Methods to collect benchmark data ideally should be cost-efficient and should result in high-quality and unbiased data. Both manual and automatic approaches have been applied. Manual curation of data can be done by experts or researchers, or through crowd workers, which comes with its own set of considerations. Automatic approaches often involve automatically generating data by applying language models or automatically extracting or mining data from existing resources. As detailed in Appendix A, a benchmark is often created through a combination of these approaches. For the rest of this section, we summarize pros and cons for these different approaches.

Manual versus automatic generation. Until recently, many of the existing benchmarks were created manually by groups of experts. This may involve tedious processes like manually collecting data from other corpora or the Internet, e.g., the first RTE Challenge (Dagan et al., 2005), or authoring most data from scratch, e.g., the Winograd Schema Challenge (Levesque, 2011; Levesque et al., 2012; Morgenstern & Ortiz, 2015; Morgenstern et al., 2016). This approach ensures data is high quality and thus typically requires little validation, however it is not scalable. These datasets are often small compared to those created with other approaches.

Recent advances in NLP make it possible to automatically generate textual data (e.g., natural language statements, questions, etc.) for benchmark tasks. Although this approach is scalable and efficient, the quality of data varies, and often depends directly on the language model used. For example, in bAbI (Weston et al., 2016), as agents interacted with objects in a virtual world and with each other, examples were automatically generated. This method ensures that produced data are sensible to the constraints of the physical world. However, as the questions and answers are written with simple structure, the data is easily understood by machines. Most of the dataset is solved with 100% accuracy just by baseline systems. Consequently, the bAbI tasks are often considered as toy tasks. A more sophisticated rule-based method for probabilistically generating text data which encourages more diverse language without any added biases is presented by Manning and Hudson (2018). Though automatic natural language generation methods are improving, it is likely that such approaches will still require some manual validation.

Automatic generation versus text mining. As millions of natural language texts are publicly available on the Internet and in existing datasets, it is possible to build language benchmarks from these texts by automatically mining texts and extracting sentences. This process is most successful

when the information source is created by experts and highly accurate. For example, in the CLOTH (Xie et al., 2017) benchmark, data instances were mined from fill-in-the-blank English tests created by human teachers. While other automatically generated cloze tasks like CBT (Hill et al., 2015) choose missing words mostly randomly, CLOTH is more challenging as the missing word in each example was chosen by an expert. For many other benchmarks that are built by mining less reliable or consistent sources, there is often a need for automatic or human validation or filtering, e.g., in creating SWAG (Zellers et al., 2018), which was mined in part from other corpora, such as the Large Scale Movie Description Challenge (Rohrbach et al., 2017) and the captions in ActivityNet (Heilbron et al., 2015).

Crowdsourcing considerations. Acquiring language data directly from crowd workers has become more feasible in recent years due to the growth of crowdsourcing platforms, e.g., Amazon Mechanical Turk. Crowdsourcing has enabled researchers to create larger datasets than ever before, however it comes with a set of considerations relating to task complexity, worker qualification, data validation, and cost optimization.

Task complexity. When creating a crowdsourcing task, it is important to consider the difficulty level of what crowd workers will be expected to do. When given overly complicated instructions, the non-expert crowd workers may find it difficult understanding the instructions and keeping them in mind while working on the task. Easy crowdsourcing tasks typically involve quick pass/fail validation or relabeling of data, e.g., in validating SNLI (Bowman et al., 2015). Difficult crowdsourcing tasks usually require crowd workers to write large texts, e.g., in creating ROCStories (Mostafazadeh et al., 2016), where workers wrote five-sentence stories following a fairly elaborate set of restrictions on story content to ensure stories were high-quality, focused, and well-organized. Such restrictions must be explained briefly and clearly, and it may take several pilot studies to ensure workers understand and follow the instructions correctly (Mostafazadeh et al., 2016).

Worker qualification. Regardless of task difficulty, efforts should be made to avoid workers submitting invalid data, whether they are trolls or unable to follow directions. This may be done through some sort of qualification task, perhaps requiring a prospective worker to recognize examples of acceptable submissions (Mostafazadeh et al., 2016), or testing a prospective worker's grammar (Richardson et al., 2013). It may also be worthwhile to identify excellent workers, and reward them and/or recruit them for more work (Mostafazadeh et al., 2016). In our own experience with crowdsourcing, we have found that if a worker produces one invalid submission, all of the worker's submissions will likely be invalid, and thus the worker should be rejected and potentially banned from the task. On the other hand, if a worker produces an excellent submission, all of the worker's submissions will likely be excellent.

Data validation. Even though crowdsourced data is produced by non-experts, data can just as easily be validated by non-experts. In creating ROCStories, Mostafazadeh et al. (2016) employ several novel methods of crowdsourced validation of crowdsourced data. Involved validation is especially necessary for difficult crowdsourcing tasks like the authoring of ROCStories, which required crowd workers to write long texts following strict guidelines. Crowdsourced data validation typically just requires a separate group of workers to review generated data and identify any bad examples, e.g., the validation of DNC benchmark data (Poliak et al., 2018a). For labeling tasks, multiple crowd workers can label the same examples, and agreement can be measured from this to estimate data quality, e.g., in creating the JOCI benchmark (Zhang et al., 2017).

For complex writing tasks where both data authoring and validation are highly involved, it may be advantageous to present crowdsourced tasks to two interacting workers at once. Reddy et al. (2018), for example, do this to create CoQA from actual human conversations about provided passages on Amazon Mechanical Turk, and achieve high data quality with minimal validation. In creating the data, the two interacting workers validate each others' work, and can even report workers who do not follow instructions, reducing the burden of worker qualification.

Cost optimization. Though hiring crowd workers is typically cheaper than hiring permanent workers, the cost may still be limiting, especially if the creator wishes to properly evaluate the quality of generated data through verification by even more crowd workers. For example, ROCStories (Mostafazadeh et al., 2016), consisting of about 50,000 well-evaluated five-sentence stories and 13,500 test cases, cost an average of 26 cents per story and an extra 10 cents per test case, resulting in a total cost near 15,000 USD to generate the dataset. If the cost of such thorough validation is an issue, validating a random sample of produced data, e.g., in validating SNLI (Bowman et al., 2015), can serve as an indicator of the overall quality of benchmark data.

Ultimately, each method of data collection has its own advantages and drawbacks. While manual authoring results in high-quality, expert-verified data, it is slow and unscalable. Meanwhile, the quality of automatically authored data is highly dependent on the language model used, and though it is faster, it may require manual verification. If using text mining rather than generating from scratch, data is more likely to be representative of human language, however manual verification may still be necessary depending on the source which data are extracted from. And lastly, crowdsourcing is a quick and convenient way to collect human-authored data following any set of criteria or restrictions, but it comes with special considerations that address the difficulty of work, qualification of workers, data validation, and cost optimization. When developing a new benchmark, the above trade-offs will need to be carefully considered.

2.2.5 Benchmark Complexity

Given the large variety of available benchmarks for natural language inference, some understanding of the complexity or difficulty of these benchmarks would be helpful. However, as demonstrated earlier in this section, there are many variations among these benchmarks. Consequently, it is not feasible to have a fair and direct comparison. As such, we do not attempt to address whether one benchmark is more complex than another. But rather we discuss various dimensions that may affect the perception of the complexity of a benchmark, and consider their possible implications.[10]

Perception of complexity. One natural way to think about the complexity of a task is to consider the performance of both machines and humans on the task; low performance may suggest high complexity, and vice versa. Figure 11 shows human performance and state-of-the-art machine performance for selected benchmarks. This figure seems to suggest that bAbI (Weston et al., 2016) is easiest for both the human and the machine. This may not be surprising, as the bAbI benchmark data is synthetically generated with simple structure, and these simpler language patterns can be well captured by deep learning models. For the majority of benchmarks, human performance is

---

10. In this analysis, data was collected from papers referenced in Section 2.1 and leaderboards linked to in Section 4 at time of writing (October 2019). Additional data collected from the leaderboards for AlphaNLI (found at `https://leaderboard.allenai.org/anli/submissions/public`), the Story Cloze Test (found at `https://competitions.codalab.org/competitions/15333#results`), and SocialIQA (found at `https://leaderboard.allenai.org/socialiqa/submissions/public`), which are not linked to in Section 4.
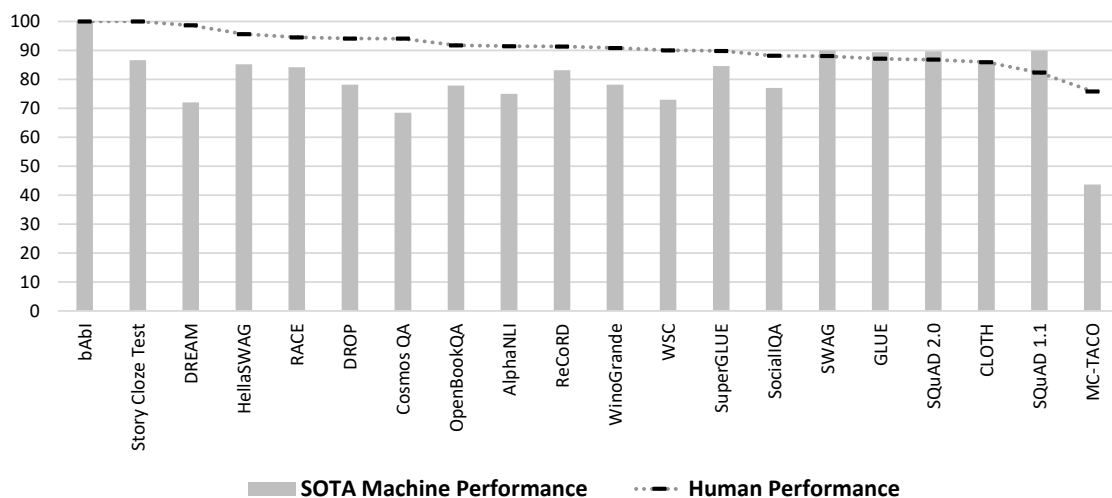
Figure 11: Comparison of state-of-the-art accuracy to human accuracy on selected benchmarks where human performance is reported.

between 80-90% accuracy. Machine performance has a much larger variance among these benchmarks. For some benchmarks, e.g., SQuAD 1.1 (Rajpurkar et al., 2016) and SWAG (Zellers et al., 2018), machine performance actually exceeds human performance. Does that mean these problems are easy and solved? On the other hand, those benchmarks where machine performance exceeds human performance have relatively low human performance. Does this suggest that the problems are actually difficult, since they seem difficult for humans? It is important to note that these performance measures, whether human or machine, may allow us to form some perception of task complexity, but they are not perfect measures for several reasons, some of which are discussed below.

Considerations for human performance measurements. Benchmark design choices, such as the evaluation framework, can affect human performance measurements. If a benchmark has high variability in possible responses, human performance is likely to be lower, but this does not necessarily mean the benchmark is extraordinarily difficult. An example of this is SQuAD 1.1, which takes a span of a passage as an answer, and has a human accuracy of 77.0% (Rajpurkar et al., 2016). While this may suggest that the task is difficult, the state-of-the-art performance has actually far exceeded human performance, suggesting the opposite. When analyzing human responses, the authors found that errors typically occurred when humans omitted unnecessary words or phrases from their answer spans. Thus, the low human performance may have largely resulted from limitations of the evaluation framework rather than the difficulty of the task.

The method of collecting human performance should also be considered. Human performance may be lower than expected if the measurement is crowdsourced, which is typical for recent benchmarks. Crowd workers want to complete tasks quickly to maximize their earnings, and thus may not spend as much time reasoning through the examples as they would in an everyday setting. Because of this, as discussed in Section 2.2.2, an additional human performance measurement collected from authors of the dataset or other experts is sometimes provided as a ceiling. However, the difference

between performance of crowd workers and experts can be high, e.g., 73.5% versus 94.5% accuracy on RACE (Lai et al., 2017). An ensemble of crowd workers may also be used to generate a higher human performance measure. Again, though, the disparity between the performance of one worker and an ensemble of workers can be high, e.g., 82.8% to 88.0% accuracy on SWAG (Zellers et al., 2018) from one worker to an ensemble of five. This uncertainty raises questions about the naturalness of these performance measurement strategies, and these factors must be considered when interpreting human performance where crowdsourcing is involved.

Considerations for machine performance measurements. Some tasks may actually be more difficult for humans than for machines based on their individual capabilities, which should be considered when comparing their performance. An example of this is CBT, whose language modeling sub-tasks require predicting random words in a sentence with or without a preceding context paragraph (Hill et al., 2015). Predicting some types of words, such as prepositions, is difficult for humans (67.6% accuracy without context), perhaps because humans excel at inferring meaning, and many prepositions have similar meanings. Consequently, neural language models, more capable of capturing patterns, easily beat humans in this sub-task of CBT. While the task is actually difficult for humans, machines are already more capable of performing it, making it less interesting of a problem to study.

On the other hand, a high machine performance does not always mean that the machine can truly perform better than humans on these tasks. As discussed in Section 2.2.3, the presence of statistical biases in benchmark data has made such a measure questionable. Further, as state-of-the-art neural approaches lack interpretability (and benchmarks do not require it), there is no way to prove this. As work progresses on eliminating these biases and providing greater explanation ability to state-of-the-art models, machine performance measurements will become more trustworthy.

Factors affecting perception of complexity. Beyond those limitations which must be considered when interpreting machine performance, some external factors can also affect the perceived complexity of the tasks. Apart from variations in the specific techniques or approaches, which will be discussed in Section 4, we find the *training data size* and the *benchmark age and popularity* most notable.

Training data size. Benchmarks with a high magnitude of training data are well-suited for deep neural networks, which can learn various patterns in language, especially in datasets with statistical biases as discussed in Section 2.2.3. Figure 12 compares overall benchmark data size with state-of-the-art accuracy on several benchmarks, including the improvement of the state of the art since each benchmark's release.

In the graph, performing a linear regression indicates no statistically significant correlation, however we can qualitatively interpret the data. With the exception of a few data points, we see a slightly positive trend in both the absolute state-of-the-art performance and the difference from the baseline performance. It appears that the highest performance and most improvement has come in benchmarks with greater than 10,000 examples. On small benchmarks with less than 1,000 examples, such as the Winograd Schema Challenge (Davis et al., 2017), we have seen limited improvement, perhaps because neural networks struggle to converge on such small data. Conversely, on large benchmarks with over 100,000 examples, such as SQuAD 1.1 (Rajpurkar et al., 2016), SWAG (Zellers et al., 2018), and GLUE (Wang et al., 2018), we see the highest state-of-the-art performance, and some of the biggest improvements in state-of-the-art performance since release.
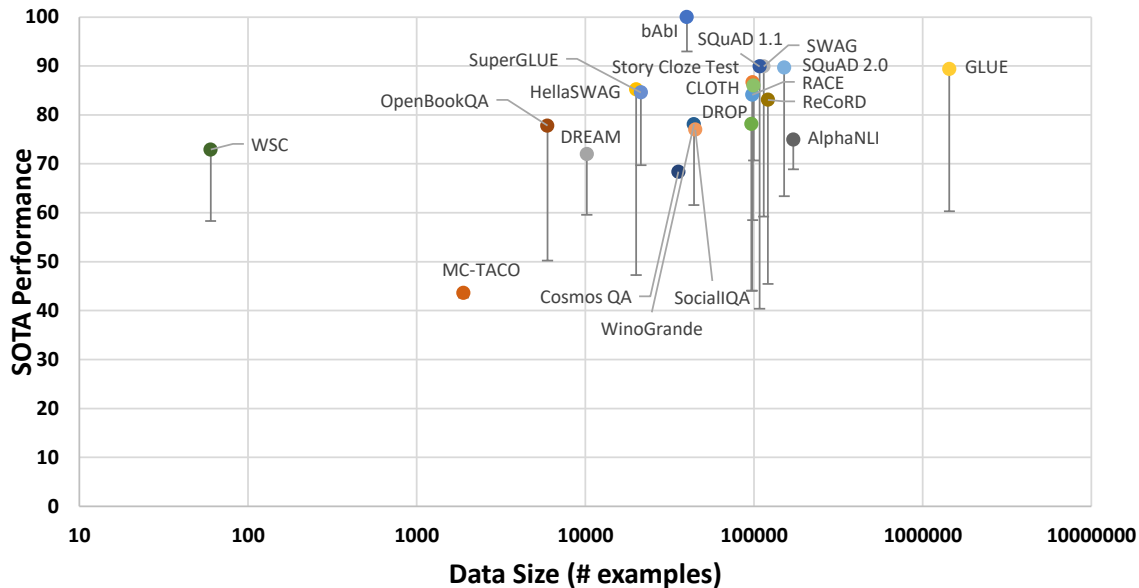
Figure 12: Data size versus state-of-the-art accuracy for selected benchmarks. Error bars indicate improvement from the original state-of-the-art accuracy reported by baselines in paper.

While this is intuitive, we do see some anomalies. For example, while bAbI (Weston et al., 2016) has a relatively small magnitude of data, state-of-the-art accuracy has reached 100%. This is likely because the data in this benchmark is synthetically generated and predictably structured, another potential factor that can make a benchmark easier. We also see large benchmarks where performance or improvement has been limited, e.g., the Story Cloze Test (Mostafazadeh et al., 2016) and Cosmos QA (Huang et al., 2019), and we predict that the age or popularity of a benchmark may factor into this.

Benchmark age. How long a benchmark has been released certainly affects how much the benchmark has been worked on, and therefore must affect the state-of-the-art machine performance. Figure 13 shows the difference between the state-of-the-art and human performance with respect to the year when the benchmark was released. While again we cannot find a statistically significant trend, many of the data points qualitatively support that we have made more progress on older benchmarks. This does not mean that the newer benchmarks are more difficult than those which we have made more progress on, rather there may have not yet been enough time or enough people working on them to achieve similar progress. However, some interesting anomalies can be identified in the graph, such as the Winograd Schema Challenge (Davis et al., 2017) and the Story Cloze Test (Mostafazadeh et al., 2016). While the former can likely be explained by very small data size, as discussed earlier, the latter has a relatively high gap from human performance despite being available for a few years and having a large magnitude of training data. This may be due to the winning entry from Schwartz et al. (2017) in a shared task showing that the benchmark has human writing style biases which enable artificially high performance on the benchmark (Mostafazadeh, Roth, Louis,
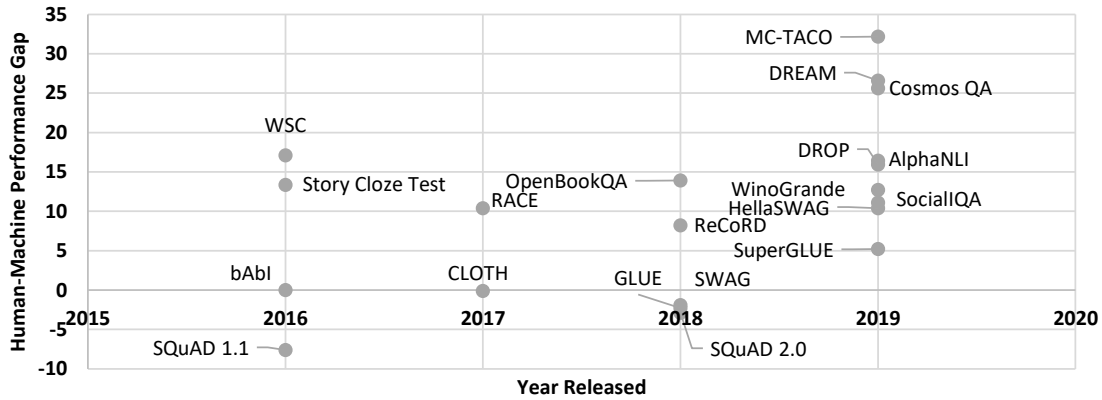
Figure 13: Year released versus gap between state-of-the-art machine accuracy and human accuracy for selected benchmarks where human performance is reported.

Chambers, & Allen, 2017), as discussed in Section 2.2.3. The last leaderboard[11] submission for this benchmark occurred over a year ago at time of writing (October 2019), so it seems researchers may have lost interest in it. Meanwhile, on other large benchmarks from the same year like bAbI (Weston et al., 2016) and SQuAD 1.1 (Rajpurkar et al., 2016), human performance has been met or exceeded.

More recent benchmarks generally have a higher gap, and this may just be because they are newer. This may also be partly explained by the recent trend of adversarial construction of data and avoiding data biases, e.g., in benchmarks like AlphaNLI (Bhagavatula et al., 2019) and Wino-Grande (Sakaguchi et al., 2019), which used adversarial methods to remove some statistical biases from the data. On the other hand, benchmarks created before this research trend are likely to be solved easily by state-of-the-art models, and thus have a lower human-machine performance gap. There are exceptions, however. For example, the Winograd Schema Challenge (Davis et al., 2017) has a higher gap from human performance than the adversarial WinoGrande (Sakaguchi et al., 2019). Again, this could be because the former benchmark does not have much data, and also because it was constructed by several experts, perhaps making it much harder for neural models to find and exploit biases in writing style.

Benchmark popularity. The popularity of a benchmark can also affect how many researchers are working on it, and consequently how much progress can be made on it and thus how difficult it may be perceived to be. If a benchmark is more popular, we may expect it to get more submissions on a leaderboard, so the gap between state-of-the-art and human performance will close rapidly. On the other hand, if a benchmark is less popular, it will receive fewer submissions, and the state-of-the-art performance will not change much, perhaps causing the benchmark to be perceived as more difficult. For selected benchmarks with online leaderboards, Figure 14 compares the number of unique, public submissions to the human-machine performance gap.

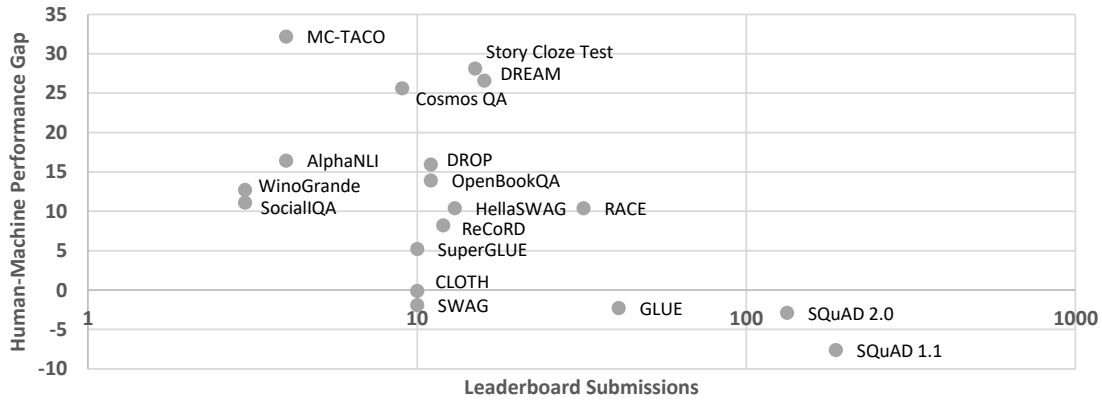11. https://competitions.codalab.org/competitions/15333#results

Figure 14: Number of leaderboard submissions versus gap between state-of-the-art machine accuracy and human accuracy for selected benchmarks where human performance is reported and public leaderboards are maintained.

For benchmarks with a low number of submissions, there is high variability, suggesting no relationship between popularity and progress on the benchmark. However, for the few benchmarks with a much higher number of submissions, we see an increasingly negative human-machine performance gap, indicating that human performance has been exceeded. The most popular benchmarks, i.e., SQuAD 1.1 (Rajpurkar et al., 2016) and 2.0 (Rajpurkar et al., 2018), have seen the most progress, with machine performance far surpassing human performance.

In conclusion, when it comes to the complexity of benchmark tasks, there is not a straightforward way to analyze this. While human and machine performance on a benchmark can shed some light on task complexity, we must consider various limitations of these performance measures themselves, as well as other factors that may hinder or drive progress in machine performance, such as benchmark data size, age, and popularity. All of these issues can affect how we perceive the complexity of benchmarks, so results must be interpreted accordingly.

3. Knowledge Resources

It is estimated that typical human adults possess up to 100,000,000 different axioms of commonsense knowledge (Chklovski, 2003). The lack of this knowledge in NLI applications is currently a major bottleneck. In order to remove this bottleneck, decades of efforts in AI have been spent on developing knowledge representations and various knowledge resources, from early works on semantic networks (Woods, 1975; Quillian, 1966) to the recent efforts of compiling large-scale commonsense knowledge graphs (Speer, Chin, & Havasi, 2017; Sap, LeBras, Allaway, Bhagavatula, Lourie, Rashkin, Roof, Smith, & Choi, 2019a). The acquired knowledge is often represented in various forms such as propositions, taxonomies, ontologies, and semantic networks. In this section, we start with an introduction to several existing knowledge resources, then discuss the main issues involved in building and leveraging these resources.

3.1 An Overview of Existing Knowledge Resources

As outlined in Section 1.2, 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* to reach a full understanding.

3.1.1 LINGUISTIC KNOWLEDGE RESOURCES

Linguistic resources have been pivotal in pushing the NLP field forward in the last thirty years. Resources have been developed where annotations for syntactic, semantic, and discourse structures are provided for training machine learning models. Several knowledge bases, particularly for lexical semantics, have also been made available to facilitate semantic processing. Also toward semantic processing, various vector embeddings of word meanings have been created. In this section, we summarize these annotated corpora, frame semantics resources, lexical resources, and pre-trained semantic vectors for linguistic knowledge.

Annotated linguistic corpora.   Widely used linguistic resources include the Penn Treebank (Marcus et al., 1993) and several derivatives of it. The Penn Treebank is perhaps the first annotated corpus that drove the development of earlier machine learning approaches in the 1990s. It started with POS tags and syntactic structures based on context-free grammar. The Penn Discourse Treebank (PDTB) is built upon it, adding annotated discourse structures (Miltsakaki, Prasad, Joshi, & Webber, 2004). OntoNotes revises information in the Wall Street Journal portion of these treebanks, integrating them with word sense, proper name, coreference, and ontological annotations, as well as including some Chinese linguistic annotations (Pradhan, Hovy, Marcus, Palmer, Ramshaw, & Weischedel, 2007). The Abstract Meaning Representation (AMR) corpus extends the Penn Treebank into a sentence-level semantic formalism (Banarescu, Bonial, Cai, Georgescu, Griffitt, Hermjakob, Knight, Koehn, Palmer, & Schneider, 2013). All of these linguistic corpora can be downloaded by members of the Linguistic Data Consortium at `https://www.ldc.upenn.edu/`.

Lexical resources.   A widely used lexical resource for commonly used nouns, verbs, adjectives and adverbs is WordNet [12] (Miller, 1995). Different from a traditional online dictionary, WordNet organizes words in terms of concepts (i.e., a list of synonyms) and their semantic relations to other words (e.g., antonymy, hyponymy/hypernymy, entailment, etc.). There are also resources specifically for verbs. VerbNet [13] by Schuler (2005) is a hierarchical English verb lexicon that is created based on the verb classes from the English Verb Classes and Alternations (EVCA) resource by Levin (1993). VerbNet defines many classes of verbs and their argument structures, selectional restrictions on the arguments, and syntactic descriptions. Other resources for verbs include VerbOcean [14](Chklovski & Pantel, 2004) which captures a network of finer-grained relations among a smaller set of common verbs, and VerbCorner [15] (Hartshorne et al., 2013) which provides crowdsourced validation for VerbNet.

Frame semantics.   Beyond annotations over large corpora or lexical hierarchies, verb semantics may also be captured by frames. Frames were defined by Minsky (1974) as cognitive data structures

---

12. `https://wordnet.princeton.edu/`
13. `https://verbs.colorado.edu/~mpalmer/projects/verbnet.html`
14. `https://demo.patrickpantel.com/demos/verbocean/`
15. `https://archive.gameswithwords.org/VerbCorner/about.php`

for information about prototypical events and situations. After further development in works by Fillmore (1976) and Fillmore (1977), the FrameNet project began annotating large texts to be used for such semantic generalization (Baker, Fillmore, & Lowe, 1998). This resulted in the FrameNet resource [16] (Fillmore, Baker, & Sato, 2002), which provides a database of semantic frames describing situations, information about them (e.g., the expected participants), relations between them, and sentences annotated for elements of the frames. It additionally includes a lexical database of nouns, verbs, and adjectives which are paired with the frames. This important resource inspired the influential and well-studied task of semantic role labeling (Gildea & Jurafsky, 2000; Punyakanok, Roth, Yih, & Zimak, 2004), where models are trained to annotate sentences for types of participants in events. PropBank [17] by Kingsbury, Palmer, and Marcus (2002) uses frames to provide the annotation of similar predicate-argument structures over the Wall Street Journal portion of the Penn Treebank (Taylor, Marcus, & Santorini, 2003). Other structures for common events have been proposed, such as plans (Miller, Galanter, & Pribram, 1960) and scripts (Schank & Abelson, 1988), which have inspired popular language tasks like the narrative cloze task (Chambers & Jurafsky, 2008).

Pre-trained semantic vectors.   Recent years have seen a surge of using continuous numeric vector representations to capture semantics of words. A detailed review of such resources is provided by Camacho-Collados and Pilehvar (2018). Early approaches evolved from first using counts of co-occurring n-grams for statistical language modeling (Brown, Della Pietra, deSouza, Lai, & Mercer, 1992), then using these counts to generate spare semantic vectors (Lund & Burgess, 1996). The most recent works, e.g., word2vec (Mikolov, Chen, Corrado, & Dean, 2013), train neural networks on word co-occurrence classification tasks generated from large texts. The learned weights from these neural networks are then used as dense vector representations of words, which can later be used as a resource to represent words in the input to neural networks for other tasks. More word embedding models have been proposed since word2vec, and the pre-trained semantic vectors from them are often made available. GloVe[18] (Pennington, Socher, & Manning, 2014) is a highly popular example, which improves efficiency and performance by training on only nonzero co-occurrence measures. Newer works have leveraged sub-word information in order to better handle rare words for which limited training data is available (Bojanowski, Grave, Joulin, & Mikolov, 2017). This can be done by splitting training text into characters when training semantic vectors, as in FastText[19] by Bojanowski et al. (2017), an extension of the word2vec models. This has also been explored by training semantic vectors for flexibly-sized pieces of words, as in the wordpiece embeddings[20] by Wu, Schuster, Chen, Le, Norouzi, Macherey, Krikun, Cao, Gao, Macherey, and others (2016).

There are many advantages to using these pre-trained semantic vectors in language processing tasks. As similar words co-occur in text and they are trained from co-occurrence, vectors for similar words appear close together in the semantic vector space. Further, they can be shown to capture word meaning to some degree. For example, in the word2vec model, it was found that when subtracting the vector for *man* from the vector for *king*, then adding the vector for *woman*, a vector approximately equal to the pre-trained vector for *queen* is produced (Mikolov et al., 2013). A major limitation of these vectors, however, is that they are context-independent, so the vector representa-

---

16. https://framenet.icsi.berkeley.edu/fndrupal/framenet_data

17. https://propbank.github.io/

18. https://nlp.stanford.edu/projects/glove/

19. https://github.com/facebookresearch/fastText

20. See https://github.com/google-research/bert for pre-trained model, or the similar SentencePiece project at https://github.com/google/sentencepiece to train a new model.

tion of a word is the same regardless of context and word sense. Most recently, contextual word representations, where representations for a word depend on where the word appears in a context, have been heavily used. More information on such representations is given in Section 4.3.3.

3.1.2 COMMON KNOWLEDGE RESOURCES

Common knowledge refers to well-known facts about the world that are often explicitly stated (Cambria et al., 2011). Since these facts are so often stated, they can be mined from the Web with relative ease to create knowledge bases. In this section, we summarize several such resources for common knowledge.

YAGO. Wikipedia is a large and open source of common knowledge. Yet Another Great Ontology (YAGO) by Suchanek, Kasneci, and Weikum (2007) augments WordNet (Miller, 1995) with common knowledge facts extracted from Wikipedia, converting WordNet from a primarily linguistic resource to a common knowledge base. YAGO originally consisted of more than 1 million entities and 5 million facts describing relationships between these entities. YAGO2 grounded entities, facts, and events in time and space, contained 446 million facts about 9.8 million entities (Hoffart, Suchanek, Berberich, & Weikum, 2012), while YAGO3 added about 1 million more entities from non-English Wikipedia articles (Mahdisoltani, Biega, & Suchanek, 2013). YAGO is available for free download at `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/`.

DBpedia. DBpedia by Auer, Bizer, Kobilarov, Lehmann, Cyganiak, and Ives (2007) is another Wikipedia-based knowledge base originally consisting of structured knowledge from more than 1.95 million Wikipedia articles. At its creation, DBpedia included around 103 million Resource Description Framework (RDF) triples[21], which are triples of subjects, predicates, and objects which describe semantic relationships. These triples included descriptions of concepts within articles, information about people, links between articles, and category labels from YAGO. The latest version, available for free at `https://wiki.dbpedia.org/develop/datasets`, consists of 6.6 million entities, 5.5 million resources classified in the DBpedia ontology, and over 23 billion RDF triples.

WikiTaxonomy. Yet another Wikipedia-based resource is WikiTaxonomy by Ponzetto and Strube (2007), which consists of about 105,000 well-evaluated semantic links between categories in Wikipedia articles. Categories and relationships are labeled using the connectivity of the conceptual network formed by the categories. The authors demonstrate that this resource can be used to calculate semantic similarity of words, which may be useful in textual entailment and other NLI tasks. WikiTaxonomy is available for free download at `https://www.h-its.org/en/research/nlp/wikitaxonomy/`.

Freebase. Freebase by Bollacker, Evans, Paritosh, Sturge, and Taylor (2008) was a knowledge graph which originally contained 125 million RDF triples of general human knowledge about 4,000 types of entities and 7,000 properties of entities. The data of Freebase was absorbed into the Google Knowledge Graph for intelligent web searching, and even migrated to Wikidata (introduced later), however the last release is still available for free download at `https://developers.google.com/freebase/`. It contains more than 1.9 billion triples.

---

21. `https://www.w3.org/TR/rdf-concepts/#section-triples`

NELL. As common knowledge is not static, it is advantageous for knowledge bases to grow over time. This is typically done through new releases, however the Never-Ending Language Learner (NELL) by Carlson, Betteridge, Kisiel, Settles, Jr, and Mitchell (2010) continually grows by automatically mining structured beliefs of varying confidence from the web daily. It originally contained 242,000 beliefs about properties of entities, but now contains over 50 million beliefs, with almost 3 million of these having high confidence. This version can be downloaded for free at `https://rtw.ml.cmu.edu/rtw/`.

Probase. Probase by Wu, Li, Wang, and Zhu (2011) is different from previous common knowledge taxonomies in that relationships are probabilistic rather than concrete. Probase consists of 2.7 million concepts extracted from 1.6 billion web pages. Relationships between concepts are described in 20.8 million is-a and is-instance-of pairs, and probabilistic interpretation is possible through provided similarity values between 0 and 1 for each pair of concepts in the knowledge base. Though the original resource is no longer available, the Microsoft Concept Graph which was built upon Probase can be downloaded for free at `https://concept.research.microsoft.com/Home/Download`.

Wikidata. Wikidata[22] is an open, collaborative, multilingual knowledge graph launched in 2012 which can be edited by anyone. At time of writing, it consisted of nearly 59 million items, which are typically commonly known entities such as famous people, artistic works, and buildings. Each item consists of various property relations linking them to other items, such as the example given on their homepage: the *highest point* property for the item *Earth* has the value *Mount Everest*. Information for using Wikidata can be found at `https://www.wikidata.org/`.

3.1.3 COMMONSENSE KNOWLEDGE RESOURCES

Commonsense knowledge, on the other hand, is considered obvious to most humans, and not likely to be explicitly stated (Cambria et al., 2011). There has been a long effort in capturing and encoding commonsense knowledge (Davis & Marcus, 2015), and various knowledge bases have been developed for this. Note that as commonsense knowledge and common knowledge are not always distinguished from each other in past work, and because some of these resources were gathered from the Web, a source of primarily common knowledge, many knowledge bases we describe here also contain common knowledge. We present several resources which focus on commonsense knowledge, but may also include common knowledge.

Cyc. A well-known project toward encoding commonsense knowledge is Cyc by Lenat and Guha (1989), a knowledge base of rules expressing ontological relationships between objects encoded in the formal logic-based CycL language. The types of objects in Cyc include entities, collections, functions, and truth functions. Cyc also includes a powerful inference engine. ResearchCyc, a release of Cyc for the research community, can be licensed for free at `https://www.cyc.com/researchcyc/`. According to this site, the latest release of ResearchCyc contains over 7 million commonsense knowledge assertions. More recently, there have been efforts to map Cyc to Wikipedia articles in an attempt to connect it to other resources such as DBpedia and Freebase (Medelyan & Legg, 2008; Pohl, 2012).

---

22. No paper has been published for Wikidata, however all information can be found at the homepage linked to at the end of this paragraph.

ConceptNet   Another popular knowledge base is ConceptNet from Liu and Singh (2004), a product of the Open Mind Common Sense project by Singh (2002), which collected free text common and commonsense knowledge from online users. This semantic network originally contained over 1.6 million assertions of commonsense knowledge represented as links between 300,000 nodes representing entities, but subsequent releases have expanded it and added more features. The latest release, ConceptNet 5.5 (Speer et al., 2017), contains over 21 million links between over 8 million nodes, having been augmented by several additional resources including Cyc (Lenat & Guha, 1989), WordNet (Fellbaum, 1999), and DBpedia (Auer et al., 2007). It includes knowledge from multilingual resources, and links to knowledge from other knowledge graphs. ConceptNet has been successfully applied in NLI systems, some of which are described in Section 4.4. ConceptNet is open; information for using or downloading it can be found at `https://conceptnet.io/`.

SenticNet.   SenticNet by Cambria, Speer, Havasi, and Hussain (2010) was originally only a commonsense knowledge base, but later versions incorporated common knowledge as well (Cambria, Olsher, & Rajagopal, 2014a). Though the knowledge base is intended for sentiment analysis, it may be useful in NLI tasks which require understanding sentiment. SenticNet is available for free at `https://sentic.net/downloads/`.

Isanette and IsaCore.   Isanette by Cambria et al. (2011) was a semantic network of both common and commonsense knowledge created by combining ProBase (Wu et al., 2011) and ConceptNet 3 (Havasi, Speer, & Alonso, 2007) into a set of "is a" relationships and confidences. This work was later cleaned and optimized into IsaCore (Cambria, Song, Wang, & Howard, 2014b), and demonstrated to be effective for sentiment analysis. IsaCore may also be a useful resource for other tasks. It is available for free download at `https://sentic.net/downloads/`.

COGBASE.   COGBASE by Olsher (2014) uses a novel formalism to represent 2.7 million concepts and 10 million commonsense facts about them. It makes up the core of SenticNet 3 (Cambria et al., 2014a). Data from COGBASE can currently be accessed via an online interface and an API available at `https://cogview.com/cogbase/`.

WebChild.   WebChild by Tandon, de Melo, Suchanek, and Weikum (2014a) was originally a commonsense knowledge base of general noun-adjective relations extracted from Web content and other resources, consisting of about 78,000 distinct noun senses, 5,600 distinct adjective senses, and 4.6 million assertions between them. These assertions captured fine-grained relations among the noun and adjective senses. Unlike other resources collected from the Web, WebChild consists primarily of commonsense knowledge, as it consists of generalized, fine-grained relationships between nouns and adjectives collected from various corpora rather than structured common knowledge taken directly from a single open source like Wikipedia. WebChild 2.0 (Tandon, de Melo, & Weikum, 2017) was later released to include over 2 million concepts and activities, and over 18 million such assertions. Data from WebChild can be browsed and downloaded online at `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/webchild/`.

LocatedNear.   Xu, Lin, and Zhu (2018a) 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 (Speer et al., 2017). They refer to this property as LocatedNear, and to address this issue, they create two datasets which we refer to jointly as LocatedNear. The first

consists of 5,000 sentences describing scenes of two objects labeled for whether the objects tend to occur near each other, which can serve as a benchmark task similar to those introduced in Section 2. The second consists of 500 pairs of objects with human-produced confidence scores for how likely the objects are to appear near each other. These resources can be downloaded from `https://github.com/adapt-sjtu/commonsense-locatednear`.

ATOMIC. The Atlas of Machine Commonsense (ATOMIC) by Sap et al. (2019a) is a crowd-sourced knowledge graph consisting of about 300,000 nodes corresponding to short textual descriptions of events, and about 877,000 "if-event-then" triples representing nine types of if-then relations between everyday events. Rather than taxonomic or ontological knowledge, this graph contains easily-accessed inferential knowledge. Sap et al. demonstrate that neural models can learn simple reasoning skills from ATOMIC which can be used to make inferences about previously unseen events. ATOMIC can be browsed and downloaded for free at `https://homes.cs.washington.edu/~msap/atomic/`.

ASER. Similar to ATOMIC, the Activities, States, Events, and Relations knowledge graph by Zhang, Liu, Pan, Song, Wing-Ki, and Leung (2019) consists of everyday events and relations between them which can be used for inference. It is on a much larger scale, though, containing 194 million events, 64 million relations, and fifteen types of relations. Further, these relations are automatically extracted from 11 billion tokens of unstructured text data rather than crowdsourced. More information about ASER and a free download link can be found at `https://hkust-knowcomp.github.io/ASER/`.

3.2 Approaches to Creating Knowledge Resources

Similar to creating the benchmarks described in Section 2, various approaches have been applied to create knowledge resources. These approaches range from manual encoding by experts, to text mining from web documents, and collection through crowdsourcing. A detailed description of these approaches is provided by Davis and Marcus (2015). Here, we give a brief discussion about pros and cons of these approaches.

Manual encoding. Early knowledge bases were often manually created. The classic example of this is Cyc, which is produced by knowledge engineers who hand-code commonsense knowledge into the CycL formalism (Lenat & Guha, 1989). Cyc has been going through continuous development over the last 35 years, since its first release in 1984. The cost of this manual encoding is high, with a total estimated cost of $120M (Paulheim, 2018). As a consequence, Cyc is small relative to other resources, and growing very slowly. On the other hand, this expert-based approach ensures high quality of data.

Text mining. Text mining and information extraction tools are often applied to automatically generate knowledge graphs and taxonomies from information sources on the Web. One popular information source is Wikipedia, which was drawn from in creating common knowledge bases such as YAGO (Suchanek et al., 2007), DBpedia (Auer et al., 2007), WikiTaxonomy (Ponzetto & Strube, 2007). Other knowledge bases are generated from crawling the Web, e.g., NELL (Carlson et al., 2010), from benchmark datasets, e.g., ATOMIC (Sap et al., 2019a), or even from other knowledge bases, e.g., IsaCore (Cambria et al., 2014b). KnowItAll (Etzioni, Cafarella, Downey, Popescu,

Shaked, Soderland, Weld, & Yates, 2005) and TextRunner (Etzioni, Banko, Soderland, & Weld, 2008) have been popular approaches for extracting this common knowledge from online sources.

However, as commonsense knowledge is rarely explicitly stated, it is more difficult to extract from text sources. Additional processes are often required to obtain knowledge beyond the text. For example, processes have been used to extract unstated inference and textual entailment rules, e.g., from Web text (Schoenmackers, Etzioni, Weld, & Davis, 2010; Gordon & Schubert, 2011), and knowledge graphs (Berant, Dagan, & Goldberger, 2011). Such rules could prove useful for many of the benchmarks introduced in Section 2. Several processes have been proposed for extracting temporal commonsense knowledge of events from text, particularly useful for benchmarks like MC-TACO (Zhou et al., 2019), which requires this knowledge. The expected durations of events can be gathered through the processes proposed by Kozareva and Hovy (2011) and Samardžić and Merlo (2016), and expected frequencies of events can be learned through work by Gordon and Schubert (2012). Information about the order of events and their participants can be extracted through processes proposed by Chambers and Jurafsky (2008) and Wang, Zhang, and Chang (2017), which may be useful in gathering commonsense knowledge about which events typically follow each other. Commonsense physical and spatial relations between objects can be inferred through processes proposed by Forbes and Choi (2017), Collell, Gool, and Moens (2018), and Yang, Birnbaum, Wang, and Downey (2018b). Relations and trends between numbers and language can also be identified and predicted through processes by Chaganty and Liang (2016) and Spithourakis and Riedel (2018). General comparative commonsense relations between entities, e.g., the knowledge that bears tend to be more dangerous than dogs, can be gathered through a process by Tandon, De Melo, and Weikum (2014b).

One key advantage of these text mining approaches is cost efficiency. According to Paulheim (2018), creating a statement in the Wikipedia-extracted DBpedia and YAGO respectively costs 1.85 cents and 0.83 cents (USD), which are hundreds of folds less than the cost of manually encoding a statement in Cyc (which was estimated at about $5.71 per statement). This makes text mining approaches easily scale up to create large knowledge bases. However, the drawback is that the acquired knowledge can be noisy and inconsistent if extracted from open online data. There exist some automatic error correction processes, e.g., a process by Spithourakis, Augenstein, and Riedel (2016) which uses language models to identify and correct textual misinterpretations of numbers. However, human validation is still likely to be required when automatically generating knowledge resources in this way (Gordon, Van Durme, & Schubert, 2010).

Crowdsourcing. Another highly popular approach to creating knowledge bases is crowdsourcing. The Open Mind Common Sense project responsible for producing ConceptNet (Liu & Singh, 2004) used a competitive online game to accept statements from humans in free text (Singh, 2002). Later, researchers converted the knowledge within collected statements into a knowledge graph by automatic processes. This method of using games to attract users to perform human intelligence tasks for free is typically referred to as Games with a Purpose (GWAP). Other GWAPs include the one used to collect annotations for VerbCorner (Hartshorne et al., 2013), and the Robot Trainer knowledge acquisition game by Rodosthenous and Michael (2016), where players must teach human knowledge to a virtual robot. The cost of crowdsourcing effort is difficult to assess. It ranges from literally getting it for free (e.g., GWAPs like Open Mind Common Sense) to an estimate of $2.25 per statement (Paulheim, 2018) for paid crowdsourcing efforts. Though the gaming approach may be cheaper in the long run, developing such a game platform is inevitably more time-consuming.

Another challenge of crowdsourcing, as pointed out by Davis and Marcus (2015), is that typical crowd workers may not be able to follow the theories and representations of knowledge that engineers have worked out. As a result, knowledge acquired by crowdsourcing can be somewhat messy, which again often needs human expert validation.

Each of these methods has its own advantages and drawbacks in terms of the trade-offs between the cost and the quality of the acquired knowledge. Most of these knowledge resources are developed from a bottom-up fashion. The goal is to create general knowledge bases to provide inductive bias for a variety of learning and reasoning tasks. Nevertheless, it is not clear whether such a goal is met and to what extent these knowledge resources are applied to natural language inference in practice. A systematic study, as suggested by Davis and Marcus (2015), for Cyc and other resources would be useful.

3.3 Leveraging Incomplete Resources

Due to the high magnitude of knowledge possessed by humans, no single knowledge resource is complete, and even the largest resources are missing essential knowledge to perform well on NLI benchmarks. This was discovered, for example, of ConceptNet (Liu & Singh, 2004), which is missing many required relations for the intuitive psychology-based Event2Mind benchmark (Rashkin et al., 2018b). To address this problem, several methods have been proposed for leveraging incomplete knowledge bases, from dimensionality reduction techniques to extractive and automatic knowledge base completion techniques.

Dimensionality reduction. One method is through dimensionality reduction. An example of this is AnalogySpace (Speer et al., 2008), an algorithm packaged with the latest release of ConceptNet (Speer et al., 2017). AnalogySpace uses principle component analysis to make analogies along dimensions such as goodness and difficulty, and smooth missing and noisy relations. Kuo and Hsu (2010) use a similar approach to compute the similarity between knowledge graph structures and draw analogies across them, producing new relations of which up to 77.6% are judged as accurate by crowd workers. The drawback of such a method is that it cannot create new terms, rather it can only generate confidence scores between any existing terms in a knowledge graph (Li, Taheri, Tu, & Gimpel, 2016).

Extractive knowledge base completion. More recently, work has aimed to complete these knowledge bases by generating new relations. Li et al. (2016) train a neural network to distinguish relations in ConceptNet from false relations, which can then be used to score the truth of proposed new relations. Further, they use underlying crowdsourced text from Open Mind Common Sense Singh (2002) to train a word embedding which maps natural language queries to rigidly-worded relations in ConceptNet. Jastrzębski, Bahdanau, Hosseini, Noukhovitch, Bengio, and Cheung (2018) use such a mapping to mine novel commonsense knowledge relations from text, while Saito, Nishida, Asano, and Tomita (2018) train a model to jointly generate and complete knowledge bases from language queries to improve accuracy of proposed relations.

Automatic knowledge base completion. Rather than extracting new relations from text, new relations may be inferred from existing relations. Angeli and Manning (2014) use a relaxed logical framework to infer novel relations from existing relations in a large knowledge base, and demonstrate that the new relations can be successfully used for inference in a textual entailment task.

More recently, neural networks are used to generate new relations. One example is memory comparison networks (Andrade, Bai, Rajendran, & Watanabe, 2018), which generalize over existing temporal relations in knowledge graphs, i.e., the typical ordering of events, in order to acquire new temporal relations. A similar recent work by Bosselut, Rashkin, Sap, Malaviya, Celikyilmaz, and Choi (2019) trains a transformer on existing relations in order to infer novel relations in knowledge graphs. When this model was trained on ConceptNet (Speer et al., 2017), human workers rated up to 92.1% of generated relations as accurate, demonstrating the potential effectiveness of this technique for already large but incomplete knowledge bases. While current efforts are promising, future work will need to continue to find and improve solutions to handle the long-tail phenomenon observed in these knowledge resources (Davis & Marcus, 2015).

4. Learning and Inference Approaches

To solve the benchmark tasks described in Section 2, a variety of approaches have been developed. These range from earlier symbolic and statistical approaches to recent approaches that apply deep learning and neural networks. This section gives a brief overview of symbolic and early statistical approaches to the surveyed benchmarks, followed by a more detailed description of representative neural approaches, which are the current state of the art on all of the benchmarks.

4.1 Symbolic Approaches

Symbolic approaches use logical forms and processes to make inferences. The study of human reasoning through logic has continued through history, from Aristotle's theories of logic and deductive reasoning (Aristotle, 1989), to the reasoning and intellect theories of philosophers like Alfarabi, Avicenna, and Averroes (Davidson, 1992). These works were a foundation for modern mathematical logic, developed in influential works like the formal logic frameworks by Morgan (1847) and Boole (1854), and the symbolic logic theories by Hilbert and Ackermann (1999) and Lewis, Langford, and Lamprecht (1959). Later work by Selman and Levesque (1990), Poole (1990), and Hobbs et al. (1993) further developed this methodology. During this evolution, Peirce (1883) proposed the process of logical abduction, i.e., the process of making a conclusion using a limited set of observations and a minimal number of assumptions, similar to plausible inference as defined by Davis and Marcus (2015) for language problems. Bayesian networks have also been used for probabilistically reasoning with logical variables (Dechter, 2013).

Meanwhile, theories of logic found applications in AI, e.g., in the early work on commonsense for machines by McCarthy (1968), and linguistics, e.g., in the theory of natural logic toward a semantic representation of language by Lakoff (1970). Fuzzy logic, which maps linguistic descriptions of numeric variables to probability distributions over the numeric variables, was an important development to handle inexactness in human language. For example, the word *young* can be used to describe someone's age, and we may specify or calculate a likelihood of this word being used to describe someone over a range of ages (Zadeh, 1975). Moore (1982) and Nunberg (1987) further motivated the role of logic in commonsense knowledge and reasoning in AI and semantic processing. By the early 1990s, symbolic approaches were dominant for knowledge representation and semantic processing of language (Birnbaum, 1991). Consequently, these approaches were heavily used for early NLI problems. A detailed review of these approaches is provided by Davis (2017).

Among the surveyed benchmarks, symbolic approaches have primarily been applied in the early RTE Challenges. One example by Raina, Ng, and Manning (2005) parsed sentences into a logical

form, then performed abduction over them using learned assumptions and likelihoods in order to determine if a high-likelihood set of assumptions can be used to prove that a sentence entails another. Giampiccolo et al. (2008) used outside semantic knowledge from resources like Wikipedia, WordNet (Miller, 1995), and VerbOcean (Chklovski & Pantel, 2004) to augment information in the hypothesis sentence, then attempted to map this to words in the premise sentence using manually authored logic rules. A later approach developed natural logic (Lakoff, 1970) into a formalism for NLI, parsing the premise and hypothesis into a natural logic form and using a decision tree to compare their features and make a decision (MacCartney & Manning, 2007). They show that combining this approach with an existing statistical RTE model exceeds the state-of-the-art performance on the third RTE Challenge dataset (Giampiccolo et al., 2007). More recently, the baseline approach to the Triangle-COPA benchmark achieved 91% accuracy by creating a comprehensive set of manually authored logic and commonsense rules to use with given mappings from the natural language in the benchmark data to logical forms (Gordon, 2016). While manually authoring logical rules and mappings from language to logical forms has been demonstrated to be highly effective in some tasks, this is not scalable for recent, larger-magnitude datasets where variation in required knowledge and language and semantic phenomena is much higher. To facilitate more practical logic-based reasoning, another body of work focuses on developing approaches to automatically mapping natural language text to a logical form (Kamath & Das, 2018), coined as *semantic parsing*. There has been an extensive amount of work on this topic, and we refer readers to the survey by Kamath and Das (2018) for a more detailed review of this work.

4.2 Early Statistical Approaches

Statistical approaches dominated the NLP field from the mid 1990s until the early 2010s. These earlier statistical approaches relied on engineered features to train various types of statistical models based on training data for a variety of early NLP benchmarks, e.g., question answering benchmarks like the classic Remedia reading comprehension dataset (Hirschman et al., 1999) and the TREC datasets (Voorhees, 2000), which were largely based on linguistic context and thus outside of the scope of this survey. Notable early approaches often combined various word matching and other lexical features with handcrafted deterministic rules Charniak, Altun, Braz, Garrett, Kosmala, Moscovich, Pang, Pyo, Sun, Wy, Yang, Zeller, and Zorn (2000), or with traditional statistical models like decision trees (Ng, Teo, & Kwan, 2000). Similar approaches were also applied to some of the benchmarks we survey here, beginning with the RTE Challenges (Dagan et al., 2005). Lexical features, for example, based on bag-of-words and word matching were commonly used in the first two RTE Challenges (Dagan et al., 2005; Bar-Haim et al., 2006), but often achieved results only slightly better than random guessing (Bar-Haim et al., 2006). More competitive systems have used more linguistic features to make predictions, such as semantic dependencies and paraphrases (Hickl, Bensley, Williams, Roberts, Rink, & Shi, 2006), synonym, antonym, and hypernym relationships derived from training data, and hidden correlation biases in benchmark data (Lai & Hockenmaier, 2014).

External knowledge and the Web were often used to complement features derived from the training data. For example, the best system in the first RTE Challenge (Dagan et al., 2005) used a naïve Bayes classifier with features from the co-occurrences of word from an online search engine (Glickman, 2006). A similar approach was also applied in the top system from the seventh RTE Challenge (Bentivogli et al., 2011), which utilized some knowledge resources from Section 3,

Task-specific output layers:

| MLP | Softmax Layer | Decoder |

Attention Mechanism

Memory Augmented

Main Architecture:

| RNN | CNN | Transformer |

Fine-tuning

Feature-based

Pre-trained language representations:

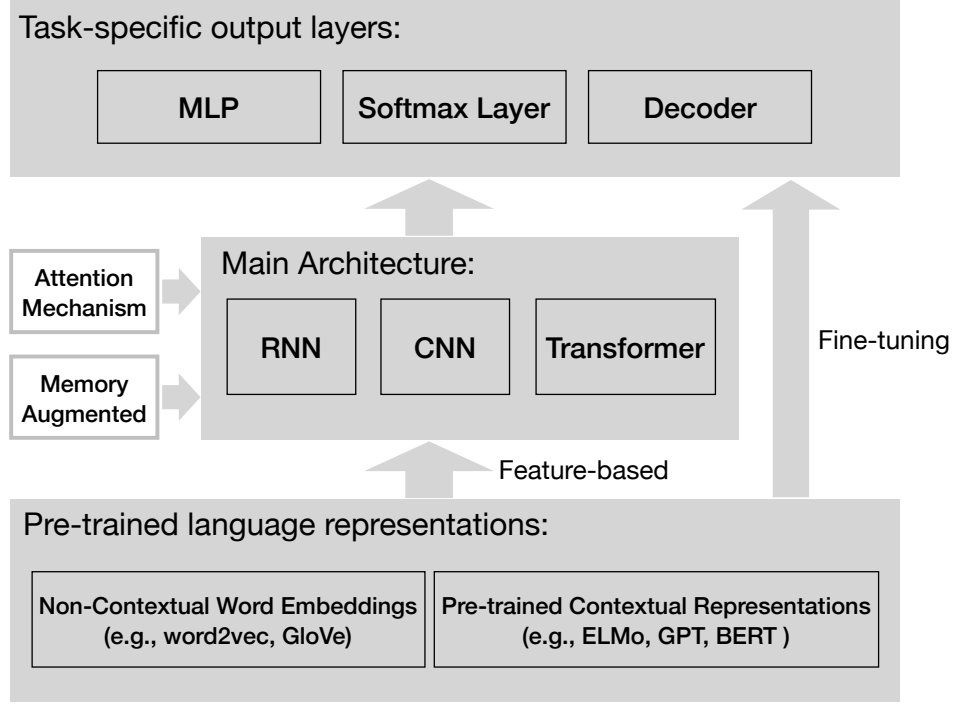| Non-Contextual Word Embeddings (e.g., word2vec, GloVe) | Pre-trained Contextual Representations (e.g., ELMo, GPT, BERT) |

Figure 15: Common components in neural approaches to natural language inference tasks.

acronyms extracted from the training data, and linguistic knowledge to calculate a statistical measure of entailment between sentences (Tsuchida & Ishikawa, 2011). While the use of some external knowledge provides an advantage over models which only use linguistic features extracted from training data, these earlier statistical models have still not been competitive in recent benchmarks of large data size. Nonetheless, such models may serve as useful baselines for new benchmarks, as demonstrated for JOCI (Zhang et al., 2017).

4.3 Neural Approaches

Neural approaches evolved from earlier statistical approaches, but use various kinds of neural network architectures to discover useful features in the data, rather than manually specifying all of the features. The increasingly large amount of data available for recent benchmarks makes it possible to train larger and deeper neural models. Today, these approaches top virtually all leaderboards for NLI benchmarks surveyed here. Figure 15 shows some common components in neural models. First of all, distributional representation of words is fundamental where word vectors or embeddings are usually trained using neural networks on large-scale text corpora. In traditional word embedding models like word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), the embedding vectors are context independent. No matter what context the target word appears in, once trained, its embedding vector is always the same. Consequently, these embeddings lack the capability of modeling different word senses in different contexts, although this phenomenon is prevalent in language. To address this problem, recent works have developed contextual word representation models, e.g., Embeddings from Language Models (ELMO) by Peters, Neumann, Iyyer, Gardner, Clark, Lee, and

Zettlemoyer (2018) and Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. (2018). ==These models give words different embedding vectors based on the context in which they appear.== These pre-trained word representations can be used as features or fine-tuned for downstream tasks. For example, the Generative Pre-trained Transformer (GPT) by Radford, Narasimhan, Salimans, and Sutskever (2018) and BERT (Devlin et al., 2018) introduce minimal task-specific parameters, and can be easily fine-tuned on the downstream tasks with modified final layers and loss functions.

On top of the word embedding layers, task-specific network architectures are designed for different downstream applications. These architectures often adopt recurrent neural networks (RNNs) such as LSTMs (Hochreiter & Schmidhuber, 1997) and GRUs (Cho, van Merrienboer, Bahdanau, & Bengio, 2014), convolutional neural networks (CNNs), or more recently, transformers (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017) to solve specific tasks. Output layers of these networks are chosen based on the task formulation. For example, a linear layer and softmax are often used for classification, while a language decoder will be used for language generation. Because of the sequential nature of language, RNN-based architectures are widely applied and are often implemented in both baseline approaches (Bowman et al., 2015; Rashkin et al., 2018b) and state-of-the-art approaches (Kim, Hong, Kang, & Kwak, 2019; Chen, Cui, Ma, Wang, Liu, & Hu, 2018; Henaff, Weston, Szlam, Bordes, & LeCun, 2017). Given different architectures, neural models also benefit from techniques like attention mechanisms and memory augmentation. For tasks that require reasoning based on multiple supporting facts, e.g., bAbI (Weston et al., 2016), memory-augmented networks like memory networks (Weston, Chopra, & Bordes, 2015) and recurrent entity networks (Henaff et al., 2017) have been shown effective. And for tasks that require alignment between input and output, e.g., textual entailment tasks like SNLI (Bowman et al., 2015), or capturing long-term dependencies, it is often beneficial to adopt attention mechanisms to models.

Next, we introduce several past and current state-of-the-art neural models for the surveyed benchmarks, particularly focusing on three aspects: attention mechanism, memory augmentation, and contextual models and representations.

4.3.1 ATTENTION MECHANISM

Since the first application of attention mechanism for neural machine translation (Bahdanau, Cho, & Bengio, 2015), attention has been used widely in NLP tasks, especially to capture the alignment between an input (encoder) and an output (decoder). Modeling attention has several advantages. It allows the decoder to directly go to and focus on certain parts of the input. It alleviates the vanishing gradient problem by providing a way to account for states far away in the input sequence. Another advantage is that the attention distribution learned by the model automatically provides an alignment between inputs and outputs which allows some understanding of their relations. Because of these advantages, attention mechanisms have been successfully applied to many NLI benchmark tasks.

Attention in RNN/CNN.   Adding an attention mechanism to RNNs, LSTMs, CNNs, and more has been shown to improve performance on various tasks compared to their vanilla models (Kim et al., 2019). ==It is particularly successful for tasks which require alignment between input and output, such as various textual entailment tasks which require the modeling of context and hypothesis==, and reading comprehension questions which refer directly to an accompanying passage, like MCScript (Ostermann et al., 2018).

For example, the official leaderboard[23] for the SNLI task (Bowman et al., 2015) reports that one of the best performing systems (Kim et al., 2019), partly inspired by DENSENET (Huang, Liu, van der Maaten, & Weinberger, 2016), uses a densely connected RNN while concatenating features from an attention mechanism to recurrent features in the network. As discussed by Kim et al. (2019), the attentive weights resulting from this alignment help the system make accurate entailment and contradiction decisions for highly similar pairs of sentences. One such example given is the context sentence "Several men in front of a white building" compared to the hypothesis sentence "Several people in front of a gray building." For the MCScript task Ostermann et al. (2018) used in SemEval 2018, online results[24] indicate that the best performer (Chen et al., 2018) achieved an accuracy of 84.13% using a bidirectional LSTM-based approach with an attention layer.

RNNs with attention also have limitations particularly when the alignment between inputs and outputs is not straightforward. For example, Chen et al. (2018) found that yes/no questions were particularly challenging in MCScript, as they require a special handling of negation and deeper understanding of the question. Further, since the multiple choices to answer questions in MCScript are human-authored instead of extracted directly from the accompanying passage like other QA benchmarks, this caused some difficulties in connecting words back to the passage that stemming alone could not fix.

Self-attention in transformers. In lieu of adding attention mechanisms to a typical neural model such as an RNN, LSTM, or CNN, the recently proposed transformer architecture is composed entirely of attention mechanisms (Vaswani et al., 2017). [25] One key difference is the self-attention layer in both the encoder and the decoder. For each word position in an input sequence, self-attention allows it to attend to all positions in the sequence to better encode the word. It provides a method to potentially capture long-range dependencies between words, such as syntactic, semantic, and coreference relations. Furthermore, instead of performing a single attention function, the transformer performs multi-head attention in the sense that it applies the attention function multiple times with different linear projections, and therefore allows the model to jointly capture different attentions from different subspaces, e.g., jointly attend to information that might indicate both coreference and syntactic relations.

Another big benefit of the transformer is its suitability for parallel computing. The sequence models such as RNN and LSTM by their sequential nature make it difficult for parallelization. The transformer, which uses attention to capture global dependencies between inputs and outputs, maximizes the amount of parallelizable computations. Empirical results on NLP tasks such as machine translation and constituency parsing have shown impressive performance gains with a significant reduction in training costs (Vaswani et al., 2017). Transformers have recently been used in pre-trained contextual models like GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) to achieve state-of-the-art performance on many of the surveyed language benchmarks.

4.3.2 MEMORY AUGMENTATION

Mentioned earlier, a popular type of approach for tasks which require comprehending passages with several state changes or supporting facts, such as bAbI (Weston et al., 2016) or ProPara (Mishra

---

23. https://nlp.stanford.edu/projects/snli/
24. https://competitions.codalab.org/competitions/17184#results
25. An excellent post on the implementation of the transformer can be found at https://nlp.seas.harvard.edu/2018/04/03/attention.html.

et al., 2018), involves augmenting systems with a dynamic memory which may be maintained over time to represent the changing state of the world. Attention mechanisms are often applied to input text to identify relevant locations or entities in memory to read or update. We discuss several recent architectures which adopt this strategy to highlight key characteristics of such an approach.

Memory networks. Memory networks by Weston et al. (2015), introduced as high-performing baseline approaches to both bAbI (Weston et al., 2016) and CBT (Hill et al., 2015), track the world state by adding a long-term memory component to the typical network architecture. A memory network consists of a memory array, an input feature map, a generalization module which updates the memory array given new input, an output feature map, and a response module which converts output to the appropriate response or action. The networks can take characters, words, or sentences as input. Each component of the network can take different forms, but a common implementation is for them to be neural networks, in which case the network is called a MEMNN. An end-to-end variant of memory networks, which can be used for a higher variety of tasks, e.g., language modeling, has also been proposed (Sukhbaatar, Szlam, Weston, & Fergus, 2015).

The ability to maintain a long-term memory provides more involved tracking of the world state and context. On the CBT cloze task (Hill et al., 2015), it is demonstrated that memory networks can outperform primarily RNN- and LSTM-based approaches in predicting missing named entities and common nouns, and this is because memory networks can efficiently leverage a wider context than these approaches in making inferences. When tested on bAbI, MEMNNs also achieved high performance and outperformed LSTM baselines, and on some tasks were able to achieve high performance with fewer training examples than provided (Weston et al., 2016).

Recurrent entity networks and variants. The recurrent entity network (ENTNET) by Henaff et al. (2017) is composed of several dynamic memory cells, where each cell learns to represent the state or properties concerning entities mentioned in the input. Each cell is a gated RNN which only updates its content when new information relevant to the particular entity is received. Further, ENTNET's memory cells run in parallel, allowing multiple locations of memory to be updated at the same time.

ENTNET is one of the first models to pass all twenty subtasks (i.e., achieve at least 95% accuracy on all twenty subtasks) in bAbI (Weston et al., 2016), achieving 99.5% accuracy over all questions. It also achieves impressive results on CBT (Hill et al., 2015), outperforming memory network baselines on both benchmarks. ENTNET is also used as a baseline in the Story Commonsense benchmark (Rashkin et al., 2018a) in an attempt to track the motivations and emotions of characters in stories from ROCStories (Mostafazadeh et al., 2016) with some success. An advantage of ENTNET is that it maintains and updates the state of the world as it reads the text, unlike memory networks, which can only perform reasoning when the entire supporting text and the question are processed and loaded to the memory. For example, given a supporting text with multiple questions, ENTNET does not need to process the input text multiple times to answer these questions, while memory networks need to re-process the whole input for each question.

While ENTNET achieved state-of-the-art performance on bAbI, it does not perform so well on ProPara (Mishra et al., 2018), another benchmark which requires tracking the world state. According to Das et al. (2019), a drawback of ENTNET is that while it maintains memory registers for entities, it has no separate embedding for individual states of entities over time. They further explain that ENTNET does not explicitly update coreferences in memory, which can certainly cause

errors when reading human-authored text which is rich in coreference, as opposed to the simply-structured, automatically-generated bAbI data.

A similar model to ENTNET is the query reduction network (QRN) by Seo, Min, Farhadi, and Hajishirzi (2017), which also uses several recurrent units as memory cells controlled by an RNN to track entities in a procedural text. Like ENTNET, the memory cells run in parallel, eliminating the vanishing gradient problem which makes it difficult for entirely RNN-based models to track long-term dependencies in procedural text. The entity representations held by the model can be queried at any point while processing the procedural text to predict states and locations of objects, not only at the end of the text as required by typical procedural text benchmarks, e.g., bAbI (Weston et al., 2016). QRNs achieve comparable results to ENTNET on bAbI.

Another similar variant, the original Differentiable Neural Computer (DNC) model proposed by Graves, Wayne, Reynolds, Harley, Danihelka, Grabska-Barwińska, Colmenarejo, Grefenstette, Ramalho, Agapiou, Badia, Hermann, Zwols, Ostrovski, Cain, King, Summerfield, Blunsom, Kavukcuoglu, and Hassabis (2016) uses several recurrent units as memory cells. Unlike the other models, DNC's memory cells are controlled by a central LSTM, whereas nearly all of ENTNET's computation occurs in the memory cells (Henaff et al., 2017). While ENTNET achieves higher performance on the bAbI benchmark (Weston et al., 2016) than the original DNC, an optimized, bidirectional form of the DNC proposed by Franke, Niehues, and Waibel (2018) exceeds the maximum performance on bAbI over all previous models, achieving a response word error rate of about 0.4. This version solves bAbI so close to perfectly that we are unlikely to see any new models attempt the benchmark.

Memory, attention, and composition networks. Similar models which may be useful toward NLI have been developed in other areas like visual question answering (VQA). One such example is the memory, attention, and composition (MAC) network (Hudson & Manning, 2018). The MAC network consists of chained memory cells each consisting of a separate memory and attentive controller. The controller attends upon the input question to update a hidden control state which determines the type of reasoning required for the question, while read and write units extract information from a knowledge source (e.g., an image or paragraph of text) guided by the controller and write it to memory. While originally applied to a VQA task, it exceeded the performance of similar network architectures which have been applied to language problems, e.g., DNC (Graves et al., 2016). As such, MAC networks may also be useful for tracking the world state in procedural text.

Neural process networks. Closely related to these models are neural process networks (NPNs) introduced by Bosselut, Levy, Holtzman, Ennis, Fox, and Choi (2018), which aim to simulate the effects of actions on entities in procedural text. It embeds input sentences in a GRU, then uses a multi-layer perceptron (MLP) to predict actions which occur in each sentence, and sentence-level and recurrent attention mechanisms to identify entities affected by actions. Selected actions and entities are fed into a simulation module which uses a recurrent unit and external knowledge of state changes caused by actions to maintain embeddings for the states of all entities in the procedural text. Unlike previous similar models, neural process networks also learn functional operators to represent actions, allowing them to explain the properties of an object that are changed by each action in a procedural text. Their results on a dataset of cooking recipes (Kiddon, Zettlemoyer, & Choi, 2016) demonstrate that learning these action embeddings allows the model to learn a more informative representation of procedural text. This representation not only allows the model to understand the text, but predict future actions in the text. Models with this suitability for learning action verb

semantics may be useful for explaining predictions on NLI benchmarks requiring intuitive physics, a capability that many current state-of-the-art models do not have.

KG-MRC. The Knowledge Graph-Machine Reading Comprehension (KG-MRC) system from Das et al. (2019), maintains a dynamic memory similar to memory networks. However, this memory is in the form of knowledge graphs generated after every sentence of procedural text, leveraging research efforts from the area of information extraction. Generated knowledge graphs are bipartite, connecting entities in the paragraph with their locations. Connections between entities and locations are updated to generate a new graph after each sentence. According to the official ProPara leaderboard[26], the Knowledge Graph-Machine Reading Comprehension (KG-MRC) system from Das et al. (2019) achieves the second-highest F-measure on the benchmark, reported as 57.60 on the leaderboard. It provides advantages over PROSTRUCT, the previous state of the art for ProPara (Tandon, Mishra, Grus, Yih, Bosselut, & Clark, 2018). While PROSTRUCT manually enforces hard and soft commonsense constraints, further investigation suggests that KG-MRC learns these constraints automatically, violating them less often than PROSTRUCT (Das et al., 2019). This shows that the use of the recurrent graph representation helps the model learn these constraints, perhaps better than can be manually enforced. Further, since KG-MRC includes a trained reading comprehension model, it can likely better track changes in coreference which often occur in these texts.

NCET. The Neural Conditional Random Field (CRF) Entity Tracking (NCET) model by Gupta and Durrett (2019) instead tracks each entity's state and location through per-entity bidirectional LSTMs. These LSTMs are trained using new, crowdsourced annotations for state and location changes of entities in the training paragraphs. A neural CRF maintains the global structure of all object states, enforcing physical laws such as their example that an object cannot move after being destroyed. According to the ProPara leaderboard, NCET is the current state of the art on the benchmark, achieving an F-measure of 62.50, a fair improvement over KG-MRC.

4.3.3 CONTEXTUAL MODELS AND REPRESENTATIONS

One of the most exciting recent advances in NLP is the development of pre-trained models and embeddings that can be used as features or further fine-tuned for downstream tasks. These models are often trained based on a large amount of unsupervised textual data. The earlier pre-trained word embedding models such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been widely applied. However, these models are context-independent, meaning the same embedding is used in different contexts, and they therefore cannot capture different word senses. More recent work has addressed this problem by pre-training models that can provide a word embedding based on context. The most representative models are ELMO, GPT, BERT and its variants, and XLNET. Next, we give a brief overview of these models and summarize their performance on the selected benchmark tasks. Table 2 quantitatively compares the performance of these models on various benchmarks.

ELMO. The characteristic contribution of Embeddings from Language Models (ELMO) is its contextual word embeddings, which each rely on the entire input sentence they belong to Peters et al. (2018). These embeddings are calculated from learned weights in a bidirectional LSTM which is pre-trained on the One Billion Word language modeling benchmark (Chelba, Mikolov, Schuster, Ge, Brants, Koehn, & Robinson, 2014). Simply adding these embeddings to the input features of

---

26. https://leaderboard.allenai.org/propara/submissions/public

previous state-of-the-art systems improved performance, suggesting that they are indeed successful in representing word context. An investigation by the authors shows that ELMO embeddings make it possible to identify word sense and POS, further supporting this.

When the ELMO embedding system was originally released, it helped exceed the state of the art on several benchmarks in the areas of question answering, textual entailment, and sentiment analysis. These included SQuAD (Rajpurkar et al., 2018) and SNLI (Bowman et al., 2015). All of these approaches have since been exceeded. ELMo still commonly appears in baseline approaches to benchmarks, e.g., to SWAG (Zellers et al., 2018) and CommonsenseQA (Talmor et al., 2019). It is often combined with the enhanced LSTM-based models such as the ESIM model (Chen, Zhu, Ling, Wei, Jiang, & Inkpen, 2017), or the CNN- and bidirectional GRU-based models such as the DocQA model (Clark & Gardner, 2018).

GPT. The Generative Pre-trained Transformer (GPT) model by Radford et al. (2018) uses the transformer architecture originally proposed by Vaswani et al. (2017), in particular the decoder. This system is pre-trained on a large amount of open online data unsupervised, then fine-tuned to various benchmark datasets. Unlike ELMO, GPT learns its contextual embeddings in an unsupervised setting, which allows it to learn features of language without restrictions. The creators found that this technique produced more discriminative features than supervised pre-training when applied to a large magnitude of clean data. The transformer architecture itself can then be easily fine-tuned in a supervised setting for downstream tasks, which ELMO is not as suitable for, and instead should be used for input features to a separate task-specific model.

When GPT was first released, it pushed the state of the art forward on 12 benchmarks in textual entailment, semantic similarity, sentiment analysis, and more. These included SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2017), SciTail (Khot et al., 2018), the Story Cloze Test (Mostafazadeh et al., 2016), COPA (Roemmele et al., 2011), and GLUE (Wang et al., 2018). GPT holds high positions on several other NLI benchmark leaderboards as well. It is often commonly used as a baseline for new benchmarks, e.g., for CommonsenseQA (Talmor et al., 2019).

Radford et al. (2018) identify several limitations of the model. First, the model has high computational requirements, which is undesirable for obvious reasons. Second, data from the Internet, which the model is pre-trained on, are incomplete and sometimes inaccurate. Lastly, like many deep learning NLP models, GPT shows some issues with generalizing over data with high lexical variation.

To improve the generalization ability and develop upon the use of unsupervised training settings, the larger GPT 2.0 was later released (Radford, Wu, Child, Luan, Amodei, & Sutskever, 2019), which is highly similar to the original implementation, but with significantly more parameters and formulated as a language model. The expanded model has achieved new state-of-the-art results on several language modeling tasks, including CBT (Hill et al., 2015), LAMBADA (Paperno et al., 2016), and the 2016 Winograd Schema Challenge (Davis et al., 2017). Further, it exceeds three out of four baseline approaches to CoQA (Reddy et al., 2018) in an unsupervised setting, i.e., trained only on documents and questions, not answers. In a supervised training setting, the model would be fed the answers directly with the questions so that model parameters could be updated based upon correlations between them. The model instead learns to perform the task by observing natural language demonstrations of it without being told where the questions and answers are. This way, it is ensured that the model is not overfitting to superficial correlations between questions and answers. A qualitative investigation into model predictions suggests that some heuristics are indeed being

learned to answer questions. For example, if asked a "who" question, the model has learned to return the name of a person mentioned in the passage that the question is posed on. This provides some evidence of the model performing genuine reasoning.

Upon reporting their results, the authors of GPT 2.0 were hesitant to release their full pre-trained model, training datasets, or training code, claiming that the model's impressive text generation ability could be used for malicious purposes such as fake news articles (Mak, 2019). However, this makes it impossible to reproduce or study their results. Critics have argued that releasing such models is important for the community to develop defenses against such attacks (Zellers, 2019). To demonstrate this, Zellers, Holtzman, Rashkin, Bisk, Farhadi, Roesner, and Choi (2019b) released a highly similar text generation model called GROVER, and showed that the best defense against it is actually itself. GROVER can be used to discriminate fake news generated by itself from real news up to 92% accuracy, while other state-of-the-art discriminative models can only achieve up to 73% accuracy. The full pre-trained models and code were finally released in November 2019,[27] but this debate has started an important discussion about the potential dangers of and defenses against the current state-of-the-art language models.

BERT. The recent Bidirectional Encoder Representations from Transformers (BERT) model proposed by Devlin et al. (2018) provides several advantages over past state-of-the-art systems with pre-trained contextual embeddings. First, it is pre-trained on larger data than previous competitive systems like GPT (Radford et al., 2018). Where GPT is pre-trained with a large text corpus, BERT is trained with two larger corpora of passages on two tasks: a cloze task where input tokens are randomly masked, and a sentence ordering task where given two sentences, the system must predict whether the second sentence could come directly after the first. This sort of transfer learning from large-scale supervised tasks has been demonstrated several times to be effective in NLP problems. Similar to GPT, BERT was further fine-tuned for each of the 11 commonsense benchmark tasks it originally attempted.

Second, BERT uses a bidirectional form of the transformer architecture (Vaswani et al., 2017) for pre-training contextual embeddings. This better captures context, an advantage that previous competitive approaches like GPT (Radford et al., 2018) and ELMo (Peters et al., 2018) do not have. Instead, GPT uses a left-to-right transformer, while ELMo uses a concatenation of left-to-right and right-to-left LSTMs.

Lastly, the structure of BERT's input embedding is advantageous for representing context, which beyond a traditional embedding for tokens, also incorporates a learned embedding for the sentence a token belongs to (i.e., in sentence pair tasks), and learned positional embeddings for tokens. Consequently, the embedding can capture each unique word and its context, perhaps in a more sophisticated way than previous systems. Further, it can flexibly represent a sentence or a pair of sentences, advantageous for solving a wide variety of language processing tasks in question answering, textual entailment, and more.

Upon its initial release, BERT exceeded the state-of-the-art accuracy on several benchmarks including GLUE (Wang et al., 2018), SQuAD 1.1 (Rajpurkar et al., 2016), and SWAG (Zellers et al., 2018) benchmarks. After this, it topped several other leaderboards, including those of CLOTH (Xie et al., 2017),[28] DREAM (Sun et al., 2019),[29] OpenBookQA (Mihaylov et al., 2018),[30]

---

27. See https://openai.com/blog/gpt-2-1-5b-release/.

28. https://www.qizhexie.com/data/CLOTH_leaderboard

29. https://dataset.org/dream/

30. https://leaderboard.allenai.org/open_book_qa/submissions/public

CoQA (Reddy et al., 2018), [31] ReCoRD (Zhang et al., 2018),[32] and SQuAD 2.0 (Rajpurkar et al., 2018). [33] Now, variants of BERT have actually topped most of the leaderboards of our surveyed benchmarks at least once, usually several times, especially for the benchmarks originally attempted by the model, e.g., GLUE. Now, BERT is seen as a baseline approach for new benchmarks, such as AlphaNLI (Bhagavatula et al., 2019) and SocialIQA (Sap et al., 2019b). In the latter, BERT was also used for a transfer learning experiment, where the model is fine-tuned to SocialIQA before fine-tuning to a new task. From this, new state-of-the-art results were achieved on COPA (Roemmele et al., 2011) and the Winograd Schema Challenge (Davis et al., 2017).

While variants of BERT are still the state of the art on many of the surveyed benchmarks, its deep, complex structure makes it unexplainable. According to Devlin et al. (2018), a goal of future work will be to determine whether BERT truly captures the intended semantic phenomena in benchmark datasets. Unfortunately, results from an early investigation by Niven and Kao (2019) suggest that like many neural models for language tasks, BERT exploits superficial correlations in data to achieve high performance, particularly on the ARCT benchmark (Habernal et al., 2018). More investigations like these for more state-of-the-art neural models on various benchmarks will be useful to gauge how much progress the models have actually made in performing genuine inference.

MT-DNN. A variant of BERT called the Multi-Task Deep Neural Network (MT-DNN) by Liu, He, Chen, and Gao (2019a) later achieved competitive performance on several leaderboards such as for SciTail [34], SNLI (Bowman et al., 2015) [35], and GLUE (Wang et al., 2018),[36] beating the original implementation of BERT. It appears that the performance gain from the MT-DNN model can mostly be attributed to adding multi-task learning during fine-tuning procedures. This is done through task-specific layers which generate representations for specific tasks, e.g., text similarity and sentence pair classification. While pre-training the bi-directional transformer helps BERT learn universal word representations which are applicable across several tasks, multi-task learning prevents the model from overfitting to particular tasks during fine-tuning, thus allowing it to leverage more cross-task data.

XLNET. A more recent model, XLNET (Yang, Dai, Yang, Carbonell, Salakhutdinov, & Le, 2019), exceeded the performance of BERT on several benchmarks upon its release, including GLUE (Wang et al., 2018), SQuAD 1.1 (Rajpurkar et al., 2016), and SQuAD 2.0 (Rajpurkar et al., 2018), and DREAM (Sun et al., 2019), exceeding human performance on the former two. This is done primarily by using a new pre-training method. A key limitation that Yang et al. identify in BERT is that since tokens are randomly masked for pre-training tasks, there is a discrepancy between language data in pre-training and fine-tuning. Further, as BERT assumes independence of these masked predicted tokens, it may miss important relations between these masked tokens, e.g., if the tokens masked are *New* and *York* in the phrase "New York" (Yang et al., 2019). To resolve this issue, Yang et al. use an autoregressive language model, which can be trained from left and right context. Instead of a bidirectional transformer, they use the newer Transformer-XL, which models long-term dependencies beyond a fixed context, as the main architecture (Dai, Yang, Yang,

---

31. https://stanfordnlp.github.io/coqa/

32. https://sheng-z.github.io/ReCoRD-explorer/

33. https://rajpurkar.github.io/SQuAD-explorer/

34. https://leaderboard.allenai.org/scitail/submissions/public

35. https://nlp.stanford.edu/projects/snli/

36. https://gluebenchmark.com/leaderboard

Carbonell, Le, & Salakhutdinov, 2019). While they lose the benefit of training on the text in both directions jointly, they propose a new permutation technique which maximizes the likelihood of a sequence of tokens and all of its permutations. This allows the model to jointly train on context from both the left and right, capturing some advantages of both an autoregressive model and a bidirectional model, and resolving the discrepancy between pre-training and fine-tuning data. Further, the permutation technique allows XLNET to capture dependencies between predicted tokens, and the authors show that the model generally captures more dependencies between tokens than previous models.

ROBERTA. The original implementation and architecture of BERT has been outperformed by several variants and other transformer-based models, some of which have been discussed earlier in this section. However, an updated version with only an optimized pre-training approach, called the Robustly Optimized BERT Approach (ROBERTA), has recently become the state of the art for some benchmarks (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, & Stoyanov, 2019b), returning to the top of the leaderboards for RACE (Lai et al., 2017),[37] SWAG (Zellers et al., 2018), and GLUE (Wang et al., 2018), exceeding human performance on the latter two. ROBERTA has also topped other leaderboards since then, such as for WinoGrande (Sakaguchi et al., 2019).[38] Some of the primary changes made to the pre-training approach include randomizing masked tokens in the cloze pre-training task for each epoch rather than keeping them the same over epochs, and adding an additional task to the next sentence prediction pre-training task, where the model must predict whether a candidate next sentence comes from the same document or not. These changes along with greater attention to hyperparameter choices allowed them to pre-train BERT to a higher potential than previously done.

ALBERT. Many of the improvements in transformer-based models come from training more parameters, but this trend becomes more difficult over time due to hardware limitations (Lan, Chen, Goodman, Gimpel, Sharma, & Soricut, 2019). A Lite BERT (ALBERT) by Lan et al. (2019) implements several novel parameter reduction techniques to increase the training speed and efficiency of BERT, allowing the model to scale up much deeper than the original large form of BERT while having fewer parameters. Further, in the sentence ordering pre-training task, they use a new self-supervised loss based upon objectives for discourse coherence. These improvements lead to superior performance on several leaderboards, including GLUE (Wang et al., 2018), SQuAD 2.0 (Rajpurkar et al., 2018), RACE (Lai et al., 2017), and DROP (Dua et al., 2019).[39]

---

37. http://www.qizhexie.com/data/RACE_leaderboard.html
38. https://leaderboard.allenai.org/winogrande/submissions/get-started
39. https://leaderboard.allenai.org/drop/submissions/public

Benchmark	Simple Baseline	ELMo	GPT	BERT	MT-DNN	XLNET	ROBERTA	ALBERT	Human
CLOTH	25.0	70.7	–	**86.0**	–	–	–	–	85.9
Cosmos QA	–	–	54.5	67.1	–	–	–	–	94.0
DREAM	33.4	59.5	55.5	66.8	–	**72.0**	–	–	95.5
GLUE	–	70.0	–	80.5	87.6	88.4	88.5	**89.4**	87.1
HellaSWAG	25.0	33.3	41.7	47.3	–	–	**85.2**		95.6
MC-TACO	17.4	26.4	–	42.7	–	–	**43.6**	–	75.8
RACE	24.9	–	59.0	72.0	–	81.8	83.2	**89.4**	94.5
SciTail	60.3	–	88.3	–	94.1	–	–	–	–
SQuAD 1.1	1.3	81.0	–	87.4	–	**89.9**	–	–	82.3
SQuAD 2.0	48.9	63.4	–	80.8	–	86.3	86.8	**89.7**	86.9
SuperGLUE	47.1	–	–	69.0	–	–	**84.6**	–	89.8
SWAG	25.0	59.1	78.0	86.3	87.1	–	**89.9**	–	88.0

Table 2: Comparison of exact-match accuracy achieved on selected benchmarks by a random or majority-choice baseline, various neural contextual embedding models, and humans. ELMO refers to the highest-performing listed approach using ELMO embeddings. Best system performance on each benchmark in bold. Information extracted from leaderboards (linked to in the first column) at time of writing (October 2019), and original papers for benchmarks introduced in Section 2.

When to fine-tune. These new pre-trained contextual models are applied to benchmark tasks in different ways. Particularly, while ELMO has been traditionally used to generate input features for a separate task-specific model, BERT-based models are typically fine-tuned on various tasks and applied to them directly. Understanding why these choices were made is important for the further development of these models.

Peters et al. (2018) investigate this difference in training the two models and compare their performance both when just extracting their output as features for another model, against when fine-tuning them to be used directly on various tasks. Their results show that ELMO's LSTM architecture can actually be fine-tuned and applied directly to downstream tasks like BERT can with some success, although it is more difficult to perform this fine-tuning on ELMO. Further, performance on sentence pair classification tasks like MultiNLI (Williams et al., 2017) and SICK (Marelli et al., 2014a) is shown to be better when the contextual embeddings generated by ELMO are instead used as input features to a separate task-specific architecture. They infer that this may be because the LSTM architecture of ELMO must consider tokens sequentially, rather than being able to compare all tokens to each other across sentence pairs like BERT's transformer architecture can. BERT's output can also be used as features for a task-specific model with some success, and it actually outperforms ELMO in most of the studied tasks when used in this way, likely for the same reason. However, they find that performance on sentence similarity tasks like the Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005) is significantly better when the model is fine-tuned to the task.

Liu, Gardner, Belinkov, Peters, and Smith (2019) further investigate the transferability of embeddings produced by these models, and find that even linear models trained with them perform comparably to state-of-the-art task-specific models. They also find that features from the lowest layer of LSTM models like ELMO are most transferable, while higher layers are more task-specific. Meanwhile, features from the middle layers of transformer models like BERT are most transferable, and embeddings are not observed to become more task-specific through layers. This may shed some more light on performance differences of these models when used for pre-trained embeddings compared to fine-tuning them on a particular task. More investigation into these contextual models will be invaluable to the research community as they continue to gain popularity.

4.4 Incorporating External Knowledge

One challenge of the current trend of work is the disconnect between knowledge resources and approaches taken to tackle those benchmark tasks. As outlined in Section 2, most of our surveyed benchmarks require a large amount of external knowledge to be solved by humans. Surprisingly, many of the recent approaches, especially neural approaches, have relied on only the benchmark training data and some linguistic resources, usually pre-trained word embeddings, to build models for reasoning and inference. Despite the availability of common and commonsense knowledge resources discussed in Section 3, none of them are actually applied to achieve state-of-the-art performance on the benchmark tasks, and only a few of them are applied in any recent approaches.

Use of linguistic resources. WordNet (Miller, 1995) is perhaps the most widely applied lexical resource, and its word relations are particularly useful for textual entailment problems. As such, WordNet has made an appearance in early statistical approaches all throughout the RTE Challenges (Dagan et al., 2005; Hickl et al., 2006; Giampiccolo et al., 2008; Iftene, 2008; Bentivogli et al., 2011; Tsuchida & Ishikawa, 2011). WordNet has also been shown to improve performance of neural models on more recent benchmarks. For example, Bauer, Wang, and Bansal (2018) use relations from WordNet to help answer questions requiring external knowledge in NarrativeQA (Kočiský et al., 2018). The Knowledge-Based Inference Model (KIM), a more involved approach by Chen, Zhu, Ling, Inkpen, and Wei (2018), improves performance on the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) benchmarks by using WordNet throughout. Pairs of words in the premise and hypothesis texts are aligned using knowledge-enriched co-attention, where if a relation exists in WordNet between two words in a pair, they have a higher attention score. Later, they use the content of these relations to help infer a class label of entailment, contradiction, or neutral. FrameNet (Fillmore et al., 2002) has also been a useful resource for external knowledge, particularly about the semantics of common events. For example, Botschen, Sorokin, and Gurevych (2018) used FrameNet to embed events recognized in the given context, concatenating this embedding to a traditional word embedding input to improve the performance of a baseline model for ARCT (Habernal et al., 2018).

Use of common knowledge resources. Popular common knowledge resources such as DBpedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007) have been used in creating benchmarks (Morgenstern et al., 2016; Choi et al., 2018), but have not been directly applied to solve any benchmarks. A possible exception is the Enhanced Language Representation with Informative Entities (ERNIE)

model, which used the common knowledge in Wikidata[40] to create knowledge-enhanced embeddings for entities mentioned in text, but was not applied to our surveyed benchmarks (Zhang, Han, Liu, Jiang, Sun, & Liu, 2019). Another related example comes from Emami, De La Cruz, Trischler, Suleman, and Cheung (2018), which uses information from search engines to aid in solving the Winograd Schema Challenge (Davis et al., 2017).

Most approaches gain relevant knowledge through benchmark training data and large pre-training texts only. While it is possible that these large pre-training texts are helpful for acquiring some linguistic and common knowledge, models likely still are lacking commonsense knowledge, which is typically unstated (Cambria et al., 2011), and thus unlikely to appear in pre-training text, which is largely acquired from the Web.

Use of commonsense knowledge resources. Cyc (Lenat & Guha, 1989) and ConceptNet (Liu & Singh, 2004) are by far the most talked-about commonsense knowledge resources available. However, Cyc does not appear in any of our surveyed approaches, while ConceptNet has only been occasionally applied. ConceptNet is used in a neural baseline approach to OpenBookQA (Mihaylov et al., 2018), where questions explicitly require outside common and commonsense knowledge. Interestingly, they find that their adding of facts from ConceptNet causes distraction which reduces performance, suggesting that the technique for selecting the appropriate relations is important to reduce distraction. Bauer et al. (2018) use a more careful technique in a knowledge-enhanced neural approach to NarrativeQA (Kočiský et al., 2018). For each concept in a question, they use ConceptNet relations to build candidate paths which span to a concept in the given story context, then to another concept in the context, then to an outside concept, which facilitates multi-step reasoning over the context and with outside knowledge. They then use mutual information and term-frequency measures to prune these paths and select the appropriate knowledge to perform this reasoning. ConceptNet can also be used to create knowledge-enhanced word embeddings. A neural model applied to COPA leverages commonsense knowledge from ConceptNet (Roemmele & Gordon, 2018) through ConceptNet-based embeddings by Li, Lee-Urban, Johnston, and Riedl (2013), which were generated by applying the word2vec skip-gram model (Mikolov et al., 2013) to commonsense knowledge tuples in ConceptNet. A recent approach to the Winograd Schema Challenge from (Liu, Jiang, Ling, Zhu, Wei, & Hu, 2017) uses a similar technique, as well as a baseline approach to SWAG (Zellers et al., 2018). Future work here will require the development of new techniques for selecting the appropriate knowledge from noisy knowledge resources, as well as the use of a greater breadth of knowledge resources.

Section 6 discusses future directions that put more emphasis on incorporating external knowledge in creating both benchmarks and models.

5. Other Related Benchmarks

While this paper intends to cover language understanding tasks for which some external knowledge or advanced reasoning beyond linguistic context is required, many related benchmarks have not been covered. First of all, nearly all language understanding benchmarks developed throughout the last couple decades could benefit from commonsense knowledge and reasoning. Second, as language communication is integral to other perception and reasoning systems, recent years have also seen an increasing number of benchmark tasks that combine language and vision.

---

40. https://www.wikidata.org/

Language-related benchmarks. Many early corpora for classical NLP tasks such as semantic role labeling, relation extraction, and paraphrase may also require commonsense knowledge and reasoning, though this was not emphasized or investigated at the time. For example, in creating the Microsoft Research Paraphrase Corpus, Dolan and Brockett (2005) found that the task of annotating text pairs was difficult to streamline because this often required commonsense, suggesting that the paraphrases within the corpus require commonsense knowledge and reasoning to identify. Some such tasks are actually included in the multi-task benchmarks like Inference is Everything (White et al., 2017), GLUE (Wang et al., 2018), and DNC (Poliak et al., 2018a).

More recent related tasks include challenging textual conversational benchmarks. One example is QuAC (Choi et al., 2018), which asks questions about Wikipedia articles through a discourse for robust contextual question answering, but does not require commonsense in the way that the similar CoQA benchmark does (Reddy et al., 2018).[41] Another related conversation dataset created by Xu, Zhou, Young, Zhao, Huang, and Zhu (2018b) explores the use of commonsense knowledge from ConceptNet in producing higher quality and more relevant responses for chatbots.

Besides English, there are also benchmarks in other languages. For example, there have been RTE datasets created in Italian[42] and Portuguese[43], and cross-lingual RTE datasets have appeared in several SemEval shared tasks over the years (Negri, Marchetti, Mehdad, Bentivogli, & Giampiccolo, 2012, 2013; Cer et al., 2017) to encourage progress in machine translation and content synchronization. There also exist various cross-lingual knowledge resources, including the latest version of ConceptNet (Speer et al., 2017), which contains relations from several multilingual resources.

Visual benchmarks. External knowledge and reasoning play an important role in integrating language and vision, for example, grounding language to perception (Gao, Doering, Yang, & Chai, 2016), language-based justification for action recognition (Yang, Gao, Sadiya, & Chai, 2018a), and visual question answering (Kafle & Kanan, 2017). Visual benchmarks requiring external knowledge include VQA benchmarks like the original VQA (Agrawal, Lu, Antol, Mitchell, Zitnick, Batra, & Parikh, 2015), other similar VQA datasets like Visual7W (Zhu, Groth, Bernstein, & Fei-Fei, 2016), and the newer GQA based on scene graphs of images (Hudson & Manning, 2019). There also exist similar datasets with synthetic images, such as CLEVR (Johnson, Hariharan, van der Maaten, Fei-Fei, Zitnick, & Girshick, 2017) and the work by Suhr, Lewis, Yeh, and Artzi (2017). They also include the tasks of commonsense action recognition and justification, which are found in the dataset by Fouhey, Kuo, Efros, and Malik (2018), and Visual Commonsense Reasoning (VCR) by Zellers, Bisk, Farhadi, and Choi (2019). Some visual datasets are geared toward specific reasoning processes, e.g., the dataset by (Winn & Muresan, 2018) for learning the directional changes associated with comparative adjectives for colors; and the verb causality dataset by Gao, Yang, Chai, and Vanderwende (2018), which provides both natural language descriptions and images for the effects of actions upon objects.

These are all image-based, but we are also beginning to see similar video-based datasets, such as Something Something (Goyal, Kahou, Michalski, Materzyńska, Westphal, Kim, Haenel, Fruend, Yianilos, Mueller-Freitag, Hoppe, Thurau, Bax, & Memisevic, 2017), which aims to evaluate visual commonsense through over 100,000 videos portraying everyday actions. We have also seen the vision-and-language navigation (VLN) task such as Room-to-Room (R2R) by Anderson, Wu,

---

41. Yatskar (2018) give a detailed account of further differences between CoQA and QuAC.
42. See https://www.evalita.it/2009/tasks/te.
43. See https://nilc.icmc.usp.br/assin/.

Teney, Bruce, Johnson, Sünderhauf, Reid, Gould, and van den Hengel (2018). Such benchmarks are important to promote progress in physically grounded inference.

6. Conclusion

The availability of data and computing resources and the rise of new learning and inference methods make this an unprecedentedly exciting time for research on natural language understanding and inference. As more benchmarks become available and performance on the benchmarks keeps growing, one central question is whether the technologies developed are in fact pushing the state-of-the-art in natural language inference. To address this question, in addition to continuous efforts on making bias-free benchmarks (as discussed in Section 2.2.3), here are a few directions we think are important to pursue in the future.

Need greater emphasis on external knowledge acquisition and incorporation.
As discussed in Section 4.4, most approaches to NLI rely on a large amount of pre-training and training data to learn the model. This is far from practical, which brings up some important questions as to how to properly incorporate external knowledge in modern approaches, and how to best acquire relevant external knowledge for the tasks at hand. This will be challenging, since as discussed in Section 3.3, most knowledge resources are incomplete, and do not contain all required information for solving the benchmarks. Beyond attempting to complete or generalize over these resources during inference as discussed in Section 3.3, another potential avenue to address the disconnect between benchmarks and knowledge resources is to jointly develop benchmark datasets and construct knowledge bases. Recently, we are seeing the creation of knowledge graphs geared toward particular benchmarks, like the ATOMIC knowledge graph by Sap et al. (2019a) which expands upon data in the Event2Mind benchmark (Rashkin et al., 2018b), and ideally provides the required relations that ConceptNet is missing for solving Event2Mind. We are also seeing the creation of benchmarks geared toward particular knowledge graphs, like CommonsenseQA (Talmor et al., 2019), where questions were drawn from subgraphs of ConceptNet (Liu & Singh, 2004), thus encouraging the use of ConceptNet in approaching the benchmark tasks. We have also seen benchmarks and knowledge resources created and released together. We have seen benchmarks being released with unstructured corpora containing some required knowledge, such as ARC (Clark et al., 2018) and OpenBookQA (Mihaylov et al., 2018). A more recent example is SherLIiC (Schmitt & Schütze, 2019), which uses typed relations extracted from Freebase (Bollacker et al., 2008) to form sentence pairs in a textual entailment problem, and is made available alongside the original relations. A tighter coupling of benchmark tasks and knowledge resources will help understand and formalize the scope of knowledge needed, and facilitate the development and evaluation of approaches that can incorporate external knowledge. Shared tasks, such as those at the Commonsense Inference in NLP workshop at EMNLP 2019,[44] have already begun to encourage the use of external knowledge in inference by providing several knowledge resources.

Since commonsense knowledge is so intuitive for humans, it is difficult for researchers to even identify and formalize the needed knowledge. It may be worthwhile to explore new task formulations beyond text that involve artificial agents (in either a simulated world or the real physical world) which can use language to communicate, to perceive, and to act. Some example tasks can be formed in the context of interactive task learning (Chai et al., 2018) or embodied question an-

---

44. `https://coinnlp.github.io/`

swering (Das, Datta, Gkioxari, Lee, Parikh, & Batra, 2018; Gordon, Kembhavi, Rastegari, Redmon, Fox, & Farhadi, 2018). Ortiz (2016) further motivates this idea, and provides suggestions of ways to implement a physically embodied Turing Test. Working with agents and observing their abilities and limitations in understanding language and grounding language to their own sensorimotor skills will allow researchers to better understand the space of commonsense knowledge from a practical standpoint, and tackle the problem of knowledge acquisition and language inference jointly.

Need greater emphasis on reasoning.

Some recent results have questioned whether state-of-the-art approaches actually perform genuine inference and reasoning for those benchmark tasks. For example, it has shown that models which exploit statistical biases in benchmark data perform poorly when the biases are removed (Niven & Kao, 2019; Zellers et al., 2019a). For another example, Jia and Liang (2017) propose an adversarial evaluation scheme for SQuAD (Rajpurkar et al., 2016) which randomly inserts distractor sentences into passages which do not change the meaning of the passage, and shows that high-performing models on the benchmark drop significantly in performance. Marasović (2018) highlights several more similar situations where modern NLI systems can break down due to small, inconsequential changes in inputs. All these findings point to an important question: how do we develop and evaluate models that can actually perform reasoning?

Davis and Marcus (2015) point out that in order for machines to perform comprehensive commonsense reasoning, as in the "piggy bank" example discussed in Section 1, there is a critical need for methods that can automatically integrate many types of reasoning such as temporal reasoning, plausible reasoning, and analogy. Thus, it is important to develop a more thorough understanding of this body of reasoning skills required in NLI. However, this is typically not given much attention in NLI benchmarks. Benchmark tasks often only evaluate the end result of inference (e.g., whether a question is correctly answered) without regard to the underlying process. Further, except for a few benchmarks, such as bAbI, which specifies sub-tasks by types of reasoning intended (Weston et al., 2016), and Event2Mind for prediction of specific psychological states surrounding events (Rashkin et al., 2018b), most benchmarks do not differentiate between types of reasoning. Useful analyses of required skills are occasionally done after benchmark creation, e.g., as done by Sugawara et al. (2017), Chu et al. (2017), and Yatskar (2018) for selected benchmarks. Nonetheless, the majority of benchmarks do not support a systematic evaluation of reasoning abilities. As such, for future benchmarks, it would be helpful to explicitly address different types of reasoning capabilities through data curation, which can provide a better understanding of machines' commonsense reasoning ability and allow for targeted improvement in different areas.

Need stronger justification and better understanding on design choices of models.

New neural models are constantly being developed. Very often, the design choices for various models are neither justified nor understood. For example, design choices like parameter tuning strategies are often overlooked in favor of more significant or interesting model improvements, but these small choices can actually have a significant effect on performance. We have seen in other AI subfields that more complex models may lead to better performance on a particular benchmark, but simpler models with better parameter tuning may later lead to comparable results. For example, in image classification, a study by Brendel and Bethge (2019) shows that nearly all improvements of recent deep neural networks over earlier bag-of-features classifiers come from better fine-tuning rather than improvements in decision processes.

NLI models are similarly vulnerable to these kinds of issues. For example, we have recently seen that the BERT model (Devlin et al., 2018), was exceeded by several variants and other transformer-based models which made major changes to loss functions (Liu et al., 2019a), model architecture, pre-training objectives (Yang et al., 2019), and model complexity, i.e., the number of parameters trained (Radford et al., 2019). Even beyond these more substantial changes to the model, smaller tweaks to various aspects of the model have resulted in hundreds of entries on leaderboards (e.g., those linked to in Section 4.3.3 and Table 2) leading only to marginal improvements. Recently, though, we saw ROBERTA exceed all of these variants and become the state of the art on several benchmarks using the same architecture as BERT, but making small changes to its pre-training tasks and more carefully selecting hyperparameters (Liu et al., 2019b). This demonstrates the importance of all model design choices, and puts into question the benefits of changes made in the earlier variants. More efforts on theoretical understanding and motivation for model design, pre-training techniques, and parameter tuning would be beneficial to better direct research efforts.

Need broader and multidimensional metrics for evaluation.
Evaluation, particularly through leaderboards, has been a critical component in NLI research. The metrics typically used on these leaderboards, e.g., accuracy and others discussed in Section 2.2.2, measure *task competence*, i.e., how well the learned model solves the benchmark task. Researchers are motivated to develop models that top the leaderboard by these kinds of performance measures. This opens up many questions about whether this competence-centric evaluation is an effective research practice. We argue that broader and multidimensional evaluation metrics should be considered in the future.

Task competence. A competence-centric evaluation, while important for pushing the state of the art, can also lead to a less productive path if not treated carefully. A significant amount of effort has been spent on tuning models or parameters to surpass other models or human performance without a good understanding of model behaviors, even on benchmarks with obvious drawbacks. For example, in Figure 14, we saw that SQuAD 1.1 (Rajpurkar et al., 2016) is clearly the most popular benchmark. It has seen hundreds of leaderboard submissions,[45] and machine performance has surpassed human performance. However, there are known issues with the benchmark. First, SQuAD 1.1 has been considered too simple, as answers to questions can be extracted directly from the text. As such, it was replaced with SQuAD 2.0 (Rajpurkar et al., 2018), a more interesting version with unanswerable questions. Further, state-of-the-art systems for SQuAD 1.1 have been shown to lack generalization to an adversarial modification of the benchmark data (Jia & Liang, 2017), suggesting that there are statistical biases in the dataset which models are overfitting to. Because of these issues, continuously improving the task competence measure on the leaderboard does not lead to anything insightful, and does not bring us closer to solving the broader task of question answering. Rather than submitting to popular leaderboards like SQuAD 1.1 hundreds of times to marginally surpass other models, researchers should be encouraged to work on some new, interesting, and/or adversarially constructed benchmarks, e.g., WinoGrande (Sakaguchi et al., 2019) and SocialIQA (Sap et al., 2019b). In addition, other evaluation metrics should also be considered in order to provide a better understanding of the progress on natural language inference, as discussed next.

---

45. https://rajpurkar.github.io/SQuAD-explorer/

Efficiency. Modern neural networks are becoming more and more complex and requiring more and more pre-training data and computational resources, with the cost to train models generally doubling every few months (Schwartz, Dodge, Smith, & Etzioni, 2019). As outlined by Rogers (2019), online leaderboards and shared tasks which rank approaches only by task competence measures exacerbate this problem, encouraging researchers with means to either use more data in pre-training or "throw a deeper network at it" to achieve state-of-the-art results. This trend discourages those researchers without the means to do this, especially academic research teams or those from emerging economies (Schwartz et al., 2019), from competing on leaderboards and in shared tasks. Factoring computational efficiency into evaluation, however, will reward approaches which make better use of available training data rather than continually augmenting pre-training data, and allow greater participation of researchers with limited computational resources. Additionally, this would encourage greener practices in AI (Schwartz et al., 2019), as training these models has a substantial carbon footprint (Strubell, Ganesh, & McCallum, 2019).

In fact, we are recently beginning to see an emphasis on model efficiency in NLI benchmarks and approaches. Some leaderboards, such as for SocialIQA (Sap et al., 2019b)[46] and AlphaNLI (Bhagavatula et al., 2019), [47] are now reporting runtime during prediction, which is a step in this direction. More efficient models for NLI are also being developed. One example is ROBERTA, where an optimized pre-training approach is developed to better utilize the parameters of the original BERT model (Devlin et al., 2018) and achieve higher performance (Liu et al., 2019b). Another example, motivated by hardware limitations, is ALBERT, where a parameter reduction technique allows the use of fewer parameters to achieve comparable results to previous models, and often better results when the model is scaled deeper (Lan et al., 2019). Lastly, DISTILBERT by Sanh, Debut, Chaumond, and Wolf (2019) compresses BERT with a small performance reduction by using model distillation, a model compression technique where a smaller model is trained to mimic the predictions of a larger model. Through this, they reduce model parameters by 40% and increase prediction speed by 60%, all while retaining 97% of its performance on GLUE (Wang et al., 2018). Efforts like the latter, purely toward model efficiency at the cost of performance, are not rewarded by the current evaluation criteria of leaderboards. Going forward, we hope such approaches will begin receiving greater recognition.

Transparency. As in many sub-fields of AI where deep learning approaches are heavily applied, transparency of the model and the ability to explain model behaviors are important. The underlying reasoning process of state-of-the-art models, e.g., BERT (Devlin et al., 2018), cannot be interpreted due to their deep, bidirectional architectures, and thus it is quite opaque why certain conclusions are made by these models. It is unclear if BERT is capturing semantic phenomena or again learning statistical biases in the NLI benchmarks, but evidence from probing studies and adversarial evaluations, e.g., the investigation by Niven and Kao (2019), is beginning to suggest that BERT is indeed learning and exploiting statistical biases in certain benchmarks. A better understanding of the behaviors of NLI systems, especially deep learning models that achieve high performance, is critical to confirm that systems are trustworthy and capable of generalization. As such, evaluating NLI systems upon their ability to explain their underlying reasoning process and more thoroughly prove their understanding could be another direction to pursue. We have seen an analogue of this in visual question answering with the VCR benchmark (Zellers et al., 2019), where models not only

---

46. https://leaderboard.allenai.org/socialiqa/submissions/public
47. https://leaderboard.allenai.org/anli/submissions/public

must answer a natural language question about an image, but must select an appropriate textual explanation for their answer from a set of choices. Although this setup may be similarly vulnerable to models overfitting on statistical artifacts of human language, it is a step in this direction.

Generalization ability. A system that can generalize should be able to perform inferences on new data and tasks with minimal training, as humans do. With better generalization, we would expect models to be able to acquire knowledge from any benchmark and apply it to other benchmarks. Domain adaptation, i.e., where the distribution of training data differs significantly from the distribution of test data, is a good step toward generalization to unseen data, and Marasović (2018) highlights recent successful approaches here. In fact, state-of-the-art models like BERT (Devlin et al., 2018) are getting somewhat closer to this capability, being pre-trained on large text corpora and capable of being fine-tuned to new problems with minimal effort or task-specific modifications. Some metrics that explicitly measure this type of adaptation and generalization will be useful. Nonetheless, the requirement of iteration over large amounts of training data in order to perform well on a benchmark is still a limitation for generalizing over multiple problems. This perhaps calls for a new practice for future NLI evaluation where no training data is provided, but rather a few examples to show what the evaluation would look like and to resolve engineering and interface issues. This may alleviate some problems caused by superficial data biases and dependence on large training data, and move toward models that can perform reasoning and generalization.

Acknowledgements

Appendix A. Comparison of Benchmark Creation Methods

The following table gives a chronological summary of methods used in creating benchmarks. It considers the steps of creating the context and/or question, the correct answer and alternatives, and any extra annotations (useful information beyond the answer which the benchmark evaluates upon), as well as validating the data. M refers to manual approaches by experts, A to automatic approaches through language generation, T to text mining, and C to crowdsourcing. Data size is included for comparison of methods used.

Benchmark (Reference)	Data Size	Context/Question	Answer	Alternatives	Annotation	Validation
RTE-1 (Dagan et al., 2005)	1.37K	M	M	–	–	M
RTE-2 (Bar-Haim et al., 2006)	1.60K	M	M	–	–	M
RTE-3 (Giampiccolo et al., 2007)	1.60K	M	M	–	–	M
RTE-4 (Giampiccolo et al., 2008)	1.00K	M	M	–	–	M
Conversational Entailment (Zhang & Chai, 2009)	875	T, M	M	–	–	M
RTE-5 (Bentivogli et al., 2009)	1.20K	M	M	–	–	M
RTE-6 (Bentivogli et al., 2010)	32.7K	T	M	–	–	M
RTE-7 (Bentivogli et al., 2011)	48.8K	T	M, T	–	–	M
COPA (Roemmele et al., 2011)	1.00K	M	M	M	–	M
MCTest (Richardson et al., 2013)	2.00K	C	C	C	–	A
SICK (Marelli et al., 2014a)	9.84K	M, T	C	–	–	C
SNLI (Bowman et al., 2015)	570K	T, C	C	–	C	C
CBT (Hill et al., 2015)	687K	T, A	A	A	–	–
Triangle-COPA (Gordon, 2016)	100	M	M	M	M	M
ROCStories (Mostafazadeh et al., 2016)	98.2K	C	C	C	–	C
WSC (Morgenstern et al., 2016)	60	M	M	–	–	M
bAbI (Weston et al., 2016)	40.0K	A	A	–	–	–
LAMBADA (Paperno et al., 2016)	10.0K	T	T	–	–	A, C
JOCI (Zhang et al., 2017)	39.1K	A, T	C	–	–	C
RACE (Lai et al., 2017)	97.7K	T	T	T	C	–
CLOTH (Xie et al., 2017)	99.4K	T	T	T	C	–
IIE (White et al., 2017)	313K	T, A	T, A	–	–	M
NarrativeQA (Kočiský et al., 2018)	46.8K	T, C	C	–	C	C
SciTail (Khot et al., 2018)	27.0K	T, C	C	–	C	C
ARC (Clark et al., 2018)	7.79K	T	T	T	–	–
MCScript (Ostermann et al., 2018)	13.9K	M, T, C	C	C	C	C
Story Commonsense (Rashkin et al., 2018a)	161K	T	C	–	C	C
Event2Mind (Rashkin et al., 2018b)	57.1K	T	C	–	–	C
ProPara (Mishra et al., 2018)	488	C	C	–	C	C
MultiRC (Khashabi et al., 2018)	9.87K	T, C	C	C	–	C
ARCT (Habernal et al., 2018)	2.45K	T,C	C	C	C	C
SQuAD 2.0 (Rajpurkar et al., 2018)	151K	T, C	C	–	–	C
CoQA (Reddy et al., 2018)	8.40K	T, C	C	–	C	C
QuAC (Choi et al., 2018)	98.4K	T, C	C	–	C	C
GLUE (Wang et al., 2018)	1.44M	T, A	T, A	–	M	–
DNC (Poliak et al., 2018a)	570K	M, A, T	M, A, T	–	–	C
OpenBookQA (Mihaylov et al., 2018)	5.96K	C	C	C	C	C
SWAG (Zellers et al., 2018)	114K	T	T, A	A	C	C
ReCoRD (Zhang et al., 2018)	121K	A, T	A	A	A	A, C
CommonsenseQA (Talmor et al., 2019)	9.40K	C	T	T	–	C
SocialIQA (Sap et al., 2019b)	44.8K	T, C	C	A, C	–	C
DREAM (Sun et al., 2019)	10.2K	T	T	T	–	–
DROP (Dua et al., 2019)	96.6K	T, C	C	–	–	C
SuperGLUE (Wang et al., 2019)	21.3K	T, A	T, A	–	C	–
HellaSWAG (Zellers et al., 2019a)	70.0K	T	T, A	A	C	C
WinoGrande (Sakaguchi et al., 2019)	44.0K	C, A	C	–	–	C, A
SherLIiC (Schmitt & Schütze, 2019)	3.99K	T, C	C	–	A	C
AlphaNLI (Bhagavatula et al., 2019)	171K	T	C, A	C, A	–	A
Cosmos QA (Huang et al., 2019)	35.6K	T, C	C	C	–	C
MC-TACO (Zhou et al., 2019)	1.89K	T, C	C	C	–	C

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015)*, Santiago, Chile. IEEE.

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., & Batra, D. (2017). VQA: Visual Question Answering. *Int. J. Comput. Vision*, *123*(1), 4–31.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., & van den Hengel, A. (2018). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA. IEEE.

Andrade, D., Bai, B., Rajendran, R., & Watanabe, Y. (2018). Leveraging knowledge bases for future prediction with memory comparison networks. *AI Communications*.

Angeli, G., & Manning, C. D. (2014). NaturalLI: Natural Logic Inference for Common Sense Reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 534–545, Doha, Qatar. Association for Computational Linguistics.

Aristotle (1989). *Prior Analytics*. Hackett Publishing.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., & Cudré-Mauroux, P. (Eds.), *Proceedings of the Semantic Web Challenge 2007 Co-Located with ISWC 2007 + ASWC 2007*, Lecture Notes in Computer Science, pp. 722–735, Busan, Korea. Springer Berlin Heidelberg.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *5*(2), 179–190.

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL '98/COLING '98)*, COLING '98, pp. 86–90, Montreal, QC, Canada. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., & Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Bauer, L., Wang, Y., & Bansal, M. (2018). Commonsense for Generative Multi-Hop Question Answering Tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Bejan, C. A., & Harabagiu, S. M. (2008). A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Belinkov, Y., Poliak, A., Shieber, S. M., Van Durme, B., & Rush, A. M. (2019). Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Belz, A., & Reiter, E. (2006). Comparing Automatic and Human Evaluation of NLG Systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2010). The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Third Text Analysis Conference (TAC 2010)*, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2011). The Seventh PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., & Magnini, B. (2009). The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, p. 15, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Berant, J., Dagan, I., & Goldberger, J. (2011). Global Learning of Typed Entailment Rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11)*, HLT '11, pp. 610–619, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W.-t., & Choi, Y. (2019). Abductive commonsense reasoning. *arXiv: 1908.05739*.

Birnbaum, L. (1991). Rigor mortis: A response to Nilsson's "Logic and artificial intelligence". *Artificial Intelligence*, *47*(1-3), 57–77.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pp. 1247–1250, New York, NY, USA. ACM.

Boole, G. (1854). *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*. Dover Publications.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., & Bjerva, J. (2017). The Groningen Meaning Bank. In *Handbook of Linguistic Annotation*, pp. 463–496. Springer.

Bosselut, A., Levy, O., Holtzman, A., Ennis, C., Fox, D., & Choi, Y. (2018). Simulating Action Dynamics with Neural Process Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Botschen, T., Sorokin, D., & Gurevych, I. (2018). Frame- and Entity-Based Knowledge for Common-Sense Argumentative Reasoning. In *Proceedings of the 5th Workshop on Argument Mining*, pp. 90–96, Brussels, Belgium. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Brendel, W., & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79–85.

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, *18*(4), 467–480.

Camacho-Collados, J., & Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *J. Artif. Int. Res.*, *63*(1), 743–788.

Cambria, E., Olsher, D., & Rajagopal, D. (2014a). SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*, Québec City, QC, Canada. AAAI Press.

Cambria, E., Song, Y., Wang, H., & Howard, N. (2014b). Semantic Multidimensional Scaling for Open- Domain Sentiment Analysis. *IEEE Intelligent Systems*, *29*(2), 44–51.

Cambria, E., Song, Y., Wang, H., & Hussain, A. (2011). Isanette: A Common and Common Sense Knowledge Base for Opinion Mining. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 315–322, Vancouver, BC, Canada. IEEE.

Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In *AAAI Fall Symposium on Commonsense Knowledge*, Menlo Park, CA, USA. AAAI Press.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr, E. R. H., & Mitchell, T. M. (2010). Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, Atlanta, GA, USA. AAAI Press.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, BC, Canada. Association for Computational Linguistics.

Chaganty, A., & Liang, P. (2016). How Much is 131 Million Dollars? Putting Numbers in Perspective with Compositional Descriptions. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 578–587, Berlin, Germany. Association for Computational Linguistics.

Chai, J. Y., Gao, Q., She, L., Yang, S., Saba-Sadiya, S., & Xu, G. (2018). Language to Action: Towards Interactive Task Learning with Physical Agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pp. 2–9, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Chambers, N., & Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, Columbus, OH, USA. Association for Computational Linguistics.

Charniak, E., Altun, Y., Braz, R. d. S., Garrett, B., Kosmala, M., Moscovich, T., Pang, L., Pyo, C., Sun, Y., Wy, W., Yang, Z., Zeller, S., & Zorn, L. (2000). Reading Comprehension Programs in a Statistical-language-processing Class. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests As Evaluation for Computer-Based Language Understanding Sytems*, ANLP/NAACL-ReadingComp '00, pp. 1–5, Seattle, WA, USA. Association for Computational Linguistics.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, Singapore, Singapore. ISCA Archive.

Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., & Wei, S. (2018). Neural Natural Language Inference Models Enhanced with External Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., & Inkpen, D. (2017). Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, BC, Canada. Association for Computational Linguistics.

Chen, Z., & Ji, H. (2009). Graph-based Event Coreference Resolution. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-4)*, pp. 54–57, Suntec, Singapore. Association for Computational Linguistics.

Chen, Z., Cui, Y., Ma, W., Wang, S., Liu, T., & Hu, G. (2018). HFL-RC System at SemEval-2018 Task 11: Hybrid Multi-Aspects Model for Commonsense Reading Comprehension. *arXiv: 1803.05655*.

Chklovski, T. (2003). Learner: A System for Acquiring Commonsense Knowledge by Analogy. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP '03)*, K-CAP '03, pp. 4–12, New York, NY, USA. ACM.

Chklovski, T., & Pantel, P. (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In Lin, D., & Wu, D. (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 33–40, Barcelona, Spain. Association for Computational Linguistics.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, Doha, Qatar. Association for Computational Linguistics.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., & Zettlemoyer, L. (2018). QuAC : Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Chu, Z., Wang, H., Gimpel, K., & McAllester, D. (2017). Broad Context Language Modeling as Reading Comprehension. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pp. 52–57, Valencia, Spain. Association for Computational Linguistics.

Clark, C., & Gardner, M. (2018). Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 845–855, Melbourne, Australia. Association for Computational Linguistics.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv: 1803.05457*.

Collell, G., Gool, L. V., & Moens, M.-F. (2018). Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, LA, USA. AAAI Press.

Dagan, I., Glickman, O., & Magnini, B. (2005). The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., & d'Alché-Buc, F. (Eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Vol. 3944, pp. 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 2978–2988, Florence, Italy. Association for Computational Linguistics.

DARPA (2018). Machine Common Sense. Broad Agency Announcement HR001119S0005, Defense Advanced Research Projects Agency, Arlington, VA, USA.

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA. IEEE.

Das, R., Munkhdalai, T., Yuan, X., Trischler, A., & McCallum, A. (2019). Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA.

Davidson, H. A. (1992). *Alfarabi, Avicenna, and Averroes on Intellect: Their Cosmologies, Theories of the Active Intellect, and Theories of Human Intellect*. Oxford University Press on Demand.

Davis, E. (2017). Logical Formalizations of Commonsense Reasoning: A Survey. *Journal of Artificial Intelligence Research*, 59, 651–723.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9), 92–103.

Davis, E., Morgenstern, L., & Ortiz, C. (2018). The Winograd Schema Challenge. https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html.

Davis, E., Morgenstern, L., & Ortiz, C. L. (2017). The First Winograd Schema Challenge at IJCAI-16. *AI Magazine; La Canada*, 38(3), 97–98.

de Marneffe, M.-C., Simons, M., & Tonhauser, J. (2019). The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn Und Bedeutung 23*, Barcelona, Spain.

Dechter, R. (2013). Reasoning with Probabilistic and Deterministic Graphical Models: Exact Algorithms. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 7(3), 1–191.

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv: 1810.04805*.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Dolan, W. B., & Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, Jeju Island, Korea.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M. (2019). DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA. Association for Computational Linguistics.

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., & Dang, H. T. (2013). SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Proceedings of the Seventh*

International Workshop on Semantic Evaluation (SemEval-2013), p. 13, Atlanta, GA, USA. Association for Computational Linguistics.

Elliott, D., & Keller, F. (2014). Comparing Automatic Evaluation Measures for Image Description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 452–457. Association for Computational Linguistics.

Emami, A., De La Cruz, N., Trischler, A., Suleman, K., & Cheung, J. C. K. (2018). A Knowledge Hunting Framework for Common Sense Reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 1949–1958, Brussels, Belgium. Association for Computational Linguistics.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, *51*(12), 68.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, *165*(1), 91–134.

Fellbaum, C. (1999). *WordNet : An Electronic Lexical Database*, Vol. 2nd printing of *Language, Speech, and Communication*. A Bradford Book, Cambridge, Mass.

Fillmore, C. J. (1976). The Need for a Frame Semantics within Linguistics. *Statistical Methods in Linguistics*, 5–29.

Fillmore, C. J. (1977). Scenes-and-Frames Semantics. *Fundamental Studies in Computer Science*, *59*, 55–82.

Fillmore, C. J., Baker, C. F., & Sato, H. (2002). The FrameNet Database and Software Tools.. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Forbes, M., & Choi, Y. (2017). Verb Physics: Relative Physical Knowledge of Actions and Objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pp. 266–276, Vancouver, Canada. Association for Computational Linguistics.

Fouhey, D. F., Kuo, W., Efros, A. A., & Malik, J. (2018). From Lifestyle VLOGs to Everyday Interactions. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA. IEEE.

Franke, J., Niehues, J., & Waibel, A. (2018). Robust and Scalable Differentiable Neural Computer for Question Answering. In *Proceedings of the Workshop on Machine Reading for Question Answering (MRQA)*, pp. 47–59, Melbourne, Australia. Association for Computational Linguistics.

Gabrilovich, E., Ringgaard, M., & Subramanya, A. (2013). FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0). *Note: http://lemurproject. org/clueweb09/FACC1/Cited by*, *5*.

Gao, Q., Doering, M., Yang, S., & Chai, J. (2016). Physical Causality of Action Verbs in Grounded Language Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1814–1824, Berlin, Germany. Association for Computational Linguistics.

Gao, Q., Yang, S., Chai, J., & Vanderwende, L. (2018). What Action Causes This? Towards Naive Physical Action-Effect Prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 934–945, Melbourne, Australia. Association for Computational Linguistics.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., & Dolan, B. (2008). The Fourth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, p. 9, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Giampiccolo, D., Magnini, B., Dagan, I., & Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pp. 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gildea, D., & Jurafsky, D. (2000). Automatic Labeling of Semantic Roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pp. 512–520, Hong Kong. Association for Computational Linguistics.

Glickman, O. (2006). *Applied Textual Entailment*. Ph.D. Thesis, Bar Ilan University.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 650–655, Melbourne, Australia. Association for Computational Linguistics.

Gordon, A. S. (2016). Commonsense Interpretation of Triangle Behavior. In *Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, AZ, USA. AAAI Press.

Gordon, D. B., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2018). IQA: Visual Question Answering in Interactive Environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4089–4098.

Gordon, J., & Schubert, L. (2011). Discovering Commonsense Entailment Rules Implicit in Sentences. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pp. 59–63, Edinburgh, Scottland, UK. Association for Computational Linguistics.

Gordon, J., & Schubert, L. K. (2012). Using Textual Patterns to Learn Expected Event Frequencies. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction*, AKBC-WEKEX '12, pp. 122–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gordon, J., Van Durme, B., & Schubert, L. (2010). Evaluation of Commonsense Knowledge with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 159–162, Los Angeles. Association for Computational Linguistics.

Goyal, R., Kahou, S. E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., & Memisevic, R. (2017). The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017)*.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M.,

Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476.

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pp. 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gupta, A., & Durrett, G. (2019). Tracking Discrete and Continuous Entity State for Process Understanding. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pp. 7–12, Minneapolis, MN, USA. Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, pp. 107–112, New Orleans, LA, USA. Association for Computational Linguistics.

Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2018). The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, pp. 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Harnad, S. (2000). Minds, Machines and Turing. *Journal of Logic, Language and Information*, *9*(4), 425–445.

Hartshorne, J. K., Bonial, C., & Palmer, M. (2013). The VerbCorner project: Toward an empirically-based semantic decomposition of verbs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pp. 1438–1442, Seattle, WA, USA. Association for Computational Linguistics.

Havasi, C., Speer, R., & Alonso, J. B. (2007). ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing (RANLP-07)*, Borovets, Bulgaria. Association for Computational Linguistics.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*, 243–259.

Heilbron, F. C., Escorcia, V., Ghanem, B., & Niebles, J. C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pp. 961–970, Boston, MA, USA. IEEE.

Henaff, M., Weston, J., Szlam, A., Bordes, A., & LeCun, Y. (2017). Tracking the World State with Recurrent Entity Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, p. 14, Palais des Congrès Neptune, Toulon, France.

Hickl, A., Bensley, J., Williams, J., Roberts, K., Rink, B., & Shi, Y. (2006). Recognizing textual entailment with LCC's GROUNDHOG system. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Hilbert, D., & Ackermann, W. (1999). *Principles of Mathematical Logic*. American Mathematical Soc.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv: 1511.02301*.

Hirschman, L., Light, M., Breck, E., & Burger, J. D. (1999). Deep Read: A Reading Comprehension System. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pp. 325–332, College Park, Maryland, USA. Association for Computational Linguistics.

Hobbs, J. R., Stickel, M. E., Appelt, D. E., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*(1), 69–142.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735–1780.

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2012). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, *194*, 28–61.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA. IEEE.

Huang, L., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, Hong Kong, China. Association for Computational Linguistics.

Hudson, D. A., & Manning, C. D. (2018). Compositional Attention Networks for Machine Reasoning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

Hudson, D. A., & Manning, C. D. (2019). GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA. IEEE.

Iftene, A. (2008). UAIC Participation at RTE4. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Iyer, S., Dandekar, N., & Csernai, K. (2017). First Quora Dataset Release: Question Pairs. https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs.

Jastrzębski, S., Bahdanau, D., Hosseini, S., Noukhovitch, M., Bengio, Y., & Cheung, J. (2018). Commonsense mining as knowledge base completion? A study on the impact of novelty. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pp. 8–16. Association for Computational Linguistics.

Jia, R., & Liang, P. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pp. 2021–2031. Association for Computational Linguistics.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. IEEE.

Kafle, K., & Kanan, C. (2017). Visual Question Answering: Datasets, Algorithms, and Future Challenges. *Computer Vision and Image Understanding*, *163*, 3–20.

Kamath, A., & Das, R. (2018). A Survey on Semantic Parsing. In *Proceedings of the 1st Conference on Automated Knowledge Base Construction (AKBC 2019)*, Amherst, MA, USA. Association for Computational Linguistics.

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., & Roth, D. (2018). 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 (NAACL HLT 2018)*, New Orleans, LA, USA. Association for Computational Linguistics.

Khot, T., Sabharwal, A., & Clark, P. (2018). SciTail: A Textual Entailment Dataset from Science Question Answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, p. 9, New Orleans, LA, USA. AAAI Press.

Kiddon, C., Zettlemoyer, L., & Choi, Y. (2016). Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 329–339, Austin, TX, USA. Association for Computational Linguistics.

Kim, S., Hong, J.-H., Kang, I., & Kwak, N. (2019). Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Inteligence (AAAI-19)*, Honolulu, HI, USA. AAAI Press.

Kingsbury, P., Palmer, M., & Marcus, M. (2002). Adding Semantic Annotation to the Penn Tree-Bank. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT '02)*, p. 5, San Diego, CA, USA. Morgan Kaufmann Publishers Inc.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, *6*, 317–328.

Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 597–606, Sydney, Australia. ACM, ACM.

Kozareva, Z., & Hovy, E. (2011). Learning Temporal Information for States and Events. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing (ICSC '11)*, ICSC '11, pp. 424–429, Washington, DC, USA. IEEE Computer Society.

Kuo, Y.-L., & Hsu, J. Y.-J. (2010). Bridging Common Sense Knowledge Bases with Analogy by Graph Similarity. In *Proceedings of the 2nd AAAI Conference on Collaboratively-Built Knowledge Sources and Artificial Intelligence*, AAAIWS'10-02, pp. 22–27, Atlanta, GA, USA. AAAI Press.

Lai, A., & Hockenmaier, J. (2014). Illinois-LH: A Denotational and Distributional Approach to Semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pp. 329–334, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods*

in Natural Language Processing (EMNLP 2017), pp. 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Lakoff, G. (1970). Linguistics and natural logic. *Synthese*, *22*(1), 151–271.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv: 1909.11942*.

Lee, K., Artzi, Y., Choi, Y., & Zettlemoyer, L. (2015). Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Lehnert, W. (1977). Human and Computational Question Answering*. *Cognitive Science*, *1*(1), 47–73.

Lenat, D. B., & Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project* (1st edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Levesque, H. J. (2011). The Winograd Schema Challenge. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford, CA, USA. AAAI Press.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge.. In *Proceedings of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning (KR2012)*, Rome, Italy. AAAI Press.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Lewis, C. I., Langford, C. H., & Lamprecht, P. (1959). *Symbolic Logic*. Dover publications New York.

Li, B., Lee-Urban, S., Johnston, G., & Riedl, M. O. (2013). Story Generation with Crowdsourced Plot Graphs. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, AAAI'13, pp. 598–604, Bellevue, Washington. AAAI Press.

Li, X., Taheri, A., Tu, L., & Gimpel, K. (2016). Commonsense Knowledge Base Completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1445–1455. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens, S. S. (Ed.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, H., & Singh, P. (2004). ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, *22*(4), 211–226.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, pp. 1073–1094. Association for Computational Linguistics.

Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., & Hu, Y. (2017). Combing Context and Commonsense Knowledge Through Neural Networks for Solving Winograd Schema Problems. In *2017 AAAI Spring Symposium Series*.

Liu, X., He, P., Chen, W., & Gao, J. (2019a). Multi-Task Deep Neural Networks for Natural Language Understanding. *arXiv: 1901.11504*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019b). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv: 1907.11692*.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

MacCartney, B., & Manning, C. D. (2007). Natural Logic for Textual Inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 193–200, Prague, Czech Republic. Association for Computational Linguistics.

Mahdisoltani, F., Biega, J., & Suchanek, F. M. (2013). YAGO3: A Knowledge Base from Multilingual Wikipedias. In *Proceedings of the 6th Biennial Conference on Innovative Data Systems Research (CIDR 2013)*, Asilomar, CA, USA.

Mak, A. (2019). When Is Technology Too Dangerous to Release to the Public?..

Manjunatha, V., Saini, N., & Davis, L. S. (2018). Explicit Bias Discovery in Visual Question Answering Models. *arXiv: 1811.07789*.

Manning, C., & Hudson, D. (2018). Towards real-world visual reasoning..

Marasović, A. (2018). NLP's generalization problem, and how researchers are tackling it..

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *1*(2).

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014a). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014b). SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland. Association for Computational Linguistics.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396.

McCarthy, J. (1968). Programs with Common Sense. In *Semantic Information Processing*, pp. 403–418. MIT Press.

McCarthy, J. (1976). An example for natural language understanding and the AI problems it raises. *Formalizing Common Sense: Papers by John McCarthy*, *355*.

Medelyan, O., & Legg, C. (2008). Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*, Chicago, IL, USA. AAAI Press.

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018). Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR 2013)*, Scottsdale, AZ, USA.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, *38*(11), 39–41.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Plans and the Structure of Behavior. Henry Holt and Co, New York, NY, US.

Miller, T., Hempelmann, C., & Gurevych, I. (2017). SemEval-2017 Task 7: Detection and Interpretation of English Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 58–68, Vancouver, Canada. Association for Computational Linguistics.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, p. 4, Lisbon, Portugal. Evaluation and Language Resources Distribution Agency.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., & van Son, C. (2016). MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Minsky, M. (1974). A Framework for Representing Knowledge. *AI Memos*.

Minsky, M. (2000). Commonsense-based interfaces. *Communications of the ACM*, *43*(8), 66–73.

Mishra, B. D., Huang, L., Tandon, N., Yih, W.-t., & Clark, P. (2018). Tracking State Changes in Procedural Text: A Challenge Dataset and Models for Process Paragraph Comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, New Orleans, LA, USA. Association for Computational Linguistics.

Moore, R. C. (1982). The Role of Logic in Knowledge Representation and Commonsense Reasoning. In *Proceedings of the Second AAAI Conference on Artificial Intelligence (AAAI '82)*, AAAI'82, pp. 428–433, Pittsburgh, PA, USA. AAAI Press.

Morgan, A. D. (1847). *Formal Logic: Or, The Calculus of Inference, Necessary and Probable*. Taylor and Walton.

Morgenstern, L. (2016). Pronoun Disambiguation Problems. http://commonsensereasoning.org/disambiguation.html.

Morgenstern, L., Davis, E., & Ortiz, C. L. (2016). Planning, Executing, and Evaluating the Winograd Schema Challenge. *AI Magazine*, *37*(1), 50–54.

Morgenstern, L., & Ortiz, C. (2015). The Winograd Schema Challenge. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, Austin, TX, USA. AAAI Press.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016). A Corpus and Cloze Evaluation Framework 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 (NAACL HLT 2016)*, San Diego, CA, USA. Association for Computational Linguistics.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., & Allen, J. (2017). LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics*, pp. 46–51, Valencia, Spain. Association for Computational Linguistics.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., & Giampiccolo, D. (2012). Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012)*, pp. 399–407, Montréal, Canada. Association for Computational Linguistics.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., & Giampiccolo, D. (2013). Semeval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval-2013)*, pp. 25–33, Atlanta, Georgia, USA. Association for Computational Linguistics.

Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., & Griffiths, T. (2018). Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Ng, H. T., Teo, L. H., & Kwan, J. L. P. (2000). A Machine Learning Approach to Answering Questions for Reading Comprehension Tests. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 124–132, Hong Kong, China. Association for Computational Linguistics.

Niven, T., & Kao, H.-Y. (2019). Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Nunberg, G. (1987). Position Paper on Common-sense and Formal Semantics. In *Proceedings of the 1987 Workshop on Theoretical Issues in Natural Language Processing (TINLAP '87)*, TINLAP '87, pp. 129–133, Las Cruces, NM, USA. Association for Computational Linguistics.

Olsher, D. (2014). Semantically-based priors and nuanced knowledge core for Big Data, Social AI, and language understanding. *Neural Networks*, *58*, 131–147.

Ortiz, C. (2016). Why We Need a Physically Embodied Turing Test and What It Might Look Like. *AI Magazine*, *37*(1), 55–62.

Ostermann, S., Modi, A., Roth, M., Thater, S., & Pinkal, M. (2018). MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., & Fernandez, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pp. 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Paulheim, H. (2018). How much is a Triple? Estimating the Cost of Knowledge Graph Creation. In *Proceedings of the 17th International Semantic Web Conference (ISWC 2018)*, Monterey, CA, USA. Springer.

Peirce, C. S. (1883). A Theory of Probable Inference. In Peirce, C. S. (Ed.), *Studies in Logic by Members of the Johns Hopkins University*, pp. 126–181. Little, Brown, and Company, Boston, MA. From the Commens Bibliography | http://www.commens.org/bibliography/collection_article/peirce-charles-s-1883-theory-probable-inference-studies-logic.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (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 (NAACL HLT 2018)*, pp. 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, pp. 1267–1273, Minneapolis, MN, USA. Association for Computational Linguistics.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pp. 3–31. Academic Press.

Pohl, A. (2012). Classifying the Wikipedia Articles into the OpenCyc Taxonomy. In *Proceedings of the Web of Linked Entities Workshop in Conjunction with the 11th International Semantic Web Conference (ISWC 2012)*, Boston, MA, USA.

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., & Van Durme, B. (2018a). Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Brussels, Belgium. Association for Computational Linguistics.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018b). Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on*

Lexical and Computational Semantics, pp. 180–191, New Orleans, LA, USA. Association for Computational Linguistics.

Ponzetto, S. P., & Strube, M. (2007). Deriving a Large Scale Taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, AAAI'07, pp. 1440–1445, Vancouver, BC, Canada. AAAI Press.

Poole, D. (1990). A methodology for using a default and abductive reasoning system. *International Journal of Intelligent Systems*, *5*(5), 521–548.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007). OntoNotes: A Unified Relational Semantic Representation. In *International Conference on Semantic Computing (ICSC 2007)*, pp. 517–526.

Punyakanok, V., Roth, D., Yih, W.-t., & Zimak, D. (2004). Semantic Role Labeling via Integer Linear Programming Inference. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, COLING '04, Geneva, Switzerland. Association for Computational Linguistics.

Quillian, M. R. (1966). Semantic Memory. Tech. rep. 2, Bolt Beranek and Newman Inc., Cambridge, MA, USA.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding with Unsupervised Learning..

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Rahman, A., & Ng, V. (2012). Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pp. 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Raina, R., Ng, A. Y., & Manning, C. D. (2005). Robust textual inference via learning and abductive reasoning. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, pp. 1099–1105, Pittsburgh, PA, USA. AAAI Press.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, TX, USA. Association for Computational Linguistics.

Rashkin, H., Bosselut, A., Sap, M., Knight, K., & Choi, Y. (2018a). Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia. Association for Computational Linguistics.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., & Choi, Y. (2018b). Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of the 56th Annual Meeting of the*

Association for Computational Linguistics (ACL 2018), Melbourne, Australia. Association for Computational Linguistics.

Reddy, S., Chen, D., & Manning, C. D. (2018). CoQA: A Conversational Question Answering Challenge. *arXiv: 1808.07042*.

Reiss, S. (2004). Multifaceted Nature of Intrinsic Motivation: The Theory of 16 Basic Desires. *Review of General Psychology*, *8*(3), 179–193.

Richardson, M., Burges, C. J. C., & Renshaw, E. (2013). MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, p. 11, Seattle, WA, USA. Association for Computational Linguistics.

Rodosthenous, C., & Michael, L. (2016). A Hybrid Approach to Commonsense Knowledge Acquisition. In *Proceedings of the 8th European Starting AI Researcher Symposium*, The Hague, the Netherlands.

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011). Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, p. 6, Stanford, CA, USA.

Roemmele, M., & Gordon, A. (2018). An Encoder-decoder Approach to Predicting Causal Relations in Stories. In *Proceedings of the First Workshop on Storytelling*, pp. 50–59, New Orleans, LA, USA. Association for Computational Linguistics.

Rogers, A. (2019). How the Transformers broke NLP leaderboards..

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., & Schiele, B. (2017). Movie Description. *Int. J. Comput. Vision*, *123*(1), 94–120.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018a). Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, pp. 8–14, New Orleans, LA, USA. Association for Computational Linguistics.

Rudinger, R., White, A. S., & Van Durme, B. (2018b). Neural Models of Factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, pp. 731–744, New Orleans, LA, USA. Association for Computational Linguistics.

Saito, I., Nishida, K., Asano, H., & Tomita, J. (2018). Commonsense Knowledge Base Completion and Generation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 141–150. Association for Computational Linguistics.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). WINOGRANDE: An adversarial winograd schema challenge at scale. *arXiv: 1907.10641*.

Samardžić, T., & Merlo, P. (2016). Aspect-based Learning of Event Duration Using Parallel Corpora..

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019*, Vancouver, BC, Canada.

Sap, M., LeBras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019a). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Inteligence (AAAI-19)*, Honolulu, HI, USA. AAAI Press.

Sap, M., Rashkin, H., Chen, D., LeBras, R., & Choi, Y. (2019b). SocialIQA: Commonsense Reasoning about Social Interactions. *arXiv: 1904.09728*.

Schank, R. C., & Abelson, R. P. (1988). *Scripts, Plans, Goals, and Understanding*. Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, US.

Schmitt, M., & Schütze, H. (2019). SherLIiC: A typed event-focused lexical inference benchmark for evaluating natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 902–914, Florence, Italy. Association for Computational Linguistics.

Schoenmackers, S., Etzioni, O., Weld, D. S., & Davis, J. (2010). Learning First-order Horn Clauses from Web Text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, EMNLP '10, pp. 1088–1098, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schuler, K. K. (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Dissertation, University of Pennsylvania.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). Green AI. *arXiv: 1907.10597*.

Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., & Smith, N. A. (2017). The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language (CoNLL 2017)*, Vancouver, BC, Canada. Association for Computational Linguistics.

Selman, B., & Levesque, H. J. (1990). Abductive and Default Reasoning: A Computational Core. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI '90)*, AAAI'90, pp. 343–348, Boston, MA, USA. AAAI Press.

Seo, M., Min, S., Farhadi, A., & Hajishirzi, H. (2017). Query-Reduction Networks for Question Answering. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France.

Sharma, R., Allen, J., Bakhshandeh, O., & Mostafazadeh, N. (2018). Tackling the Story Ending Biases in The Story Cloze Test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 752–757, Melbourne, Australia. Association for Computational Linguistics.

Singh, P. (2002). The Public Acquisition of Commonsense Knowledge. In *AAAI Spring Symposium Series*, Palo Alto, CA, USA.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, p. 12, Seattle, WA, USA. Association for Computational Linguistics.

Song, H.-j., Baillargeon, R., & Fisher, C. (2005). Can infants attribute to an agent a disposition to perform a particular action?. *Cognition*, *98*(2), B45–B55.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) and the Twenty-Ninth Innovative Applications of Artificial Intelligence Conference (IAAI-17)*, San Francisco, CA, USA. AAAI Press.

Speer, R., Havasi, C., & Lieberman, H. (2008). AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*, Chicago, IL, USA.

Spithourakis, G., Augenstein, I., & Riedel, S. (2016). Numerically Grounded Language Models for Semantic Error Correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 987–992, Austin, Texas. Association for Computational Linguistics.

Spithourakis, G., & Riedel, S. (2018). Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 3645–3650, Florence, Italy. Association for Computational Linguistics.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, p. 10, Amherst, MA, USA. ACM.

Sugawara, S., Yokono, H., & Aizawa, A. (2017). Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, CA, USA. AAAI Press.

Suhr, A., Lewis, M., Yeh, J., & Artzi, Y. (2017). A Corpus of Natural Language for Visual Reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pp. 217–223. Association for Computational Linguistics.

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-To-End Memory Networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 2440–2448. Curran Associates, Inc.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., & Cardie, C. (2019). DREAM: A Challenge Dataset and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, *7*(0), 217–231.

Sundheim, B. M. (1993). Tipster/MUC-5: Information extraction system evaluation. In *Proceedings of the 5th Conference on Message Understanding*, MUC5 '93, pp. 27–44, Baltimore, MD, USA. Association for Computational Linguistics.

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA. Association for Computational Linguistics.

Tandon, N., de Melo, G., Suchanek, F., & Weikum, G. (2014a). WebChild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*, pp. 523–532, New York, NY, USA. ACM.

Tandon, N., De Melo, G., & Weikum, G. (2014b). Acquiring Comparative Commonsense Knowledge from the Web. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*, AAAI'14, pp. 166–172, Québec City, QC, Canada. AAAI Press.

Tandon, N., de Melo, G., & Weikum, G. (2017). WebChild 2.0 : Fine-Grained Commonsense Knowledge Distillation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 115–120, Vancouver, BC, Canada. Association for Computational Linguistics.

Tandon, N., Mishra, B. D., Grus, J., Yih, W.-t., Bosselut, A., & Clark, P. (2018). Reasoning about Actions and State Changes by Injecting Commonsense Knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Brussels, Belgium. Association for Computational Linguistics.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: An Overview. In *Treebanks*, Vol. 20 of *Text, Speech and Language Technology*. Springer, Dordrecht.

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Bulletin*, *30*(4), 415–433.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, CONLL '03, pp. 142–147, Edmonton, AB, Canada. Association for Computational Linguistics.

Trichelair, P., Emami, A., Trischler, A., Suleman, K., & Cheung, J. C. K. (2018). How reasonable are common-sense reasoning tasks: A case-study on the winograd schema challenge and SWAG. *arXiv: 1811.01778*.

Tsuchida, M., & Ishikawa, K. (2011). IKOMA at TAC2011: A Method for Recognizing Textual Entailment using Lexical-level and Sentence Structure-level features. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc.

Voorhees, E. M. (2000). The TREC-8 Question Answering Track. *Natural Language Engineering*, *7*, 361–378.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv: 1905.00537*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium. Association for Computational Linguistics.

Wang, Z., Zhang, Y., & Chang, C.-Y. (2017). Integrating Order Information and Event Relation for Script Event Prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pp. 57–67, Copenhagen, Denmark. Association for Computational Linguistics.

Warstadt, A., Singh, A., & Bowman, S. R. (2018). Neural Network Acceptability Judgments. *arXiv: 1805.12471*.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2016). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory Networks. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA.

White, A. S., Rastogi, P., Duh, K., & Van Durme, B. (2017). Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP 2017)*, pp. 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

White, A. S., & Rawlins, K. (2018). The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, Amherst, MA, USA. GLSA Publications.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, New Orleans, LA, USA. Association for Computational Linguistics.

Winn, O., & Muresan, S. (2018). 'Lighter' Can Still Be Dark: Modeling Comparative Color Descriptions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 790–795, Melbourne, Australia. Association for Computational Linguistics.

Winograd, T. (1972). *Understanding Natural Language*. Academic Press, New York.

Woods, W. A. (1975). What's in a Link: Foundations for Semantic Networks. Tech. rep. BBN-3072, Bolt Beranek and Newman Inc., Cambridge, MA, USA.

Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2011). Towards a Probabilistic Taxonomy of Many Concepts. http://www.microsoft.com/en-us/research/wp-content/uploads/2011/03/paper-2.pdf.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., & others (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv: 1609.08144*.

Xie, Q., Lai, G., Dai, Z., & Hovy, E. (2017). Large-scale Cloze Test Dataset Created by Teachers. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark. Association for Computational Linguistics.

Xu, F. F., Lin, B. Y., & Zhu, K. (2018a). Automatic Extraction of Commonsense LocatedNear Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 96–101, Melbourne, Australia. Association for Computational Linguistics.

Xu, J., Zhou, H., Young, T., Zhao, H., Huang, M., & Zhu, X. (2018b). Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pp. 4623–4629, Stockholm, Sweden. IJCAI.

Yang, D., Lavie, A., Dyer, C., & Hovy, E. (2015). Humor Recognition and Humor Anchor Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Yang, S., Gao, Q., Sadiya, S., & Chai, J. (2018a). Commonsense Justification for Action Explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 2627–2637, Brussels, Belgium. Association for Computational Linguistics.

Yang, Y., Birnbaum, L., Wang, J.-P., & Downey, D. (2018b). Extracting Commonsense Properties from Embeddings with Limited Human Guidance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 644–649, Melbourne, Australia. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv: 1906.08237*.

Yatskar, M. (2018). A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, Minneapolis, MN, USA. Association for Computational Linguistics.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences*, *8*(3), 199–249.

Zang, L.-J., Cao, C., Cao, Y.-N., Wu, Y.-M., & CAO, C.-G. (2013). A Survey of Commonsense Knowledge Acquisition. *Journal of Computer Science and Technology*, *28*(4), 689–719.

Zellers, R. (2019). Why We Released Grover..

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA. IEEE.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing (EMNLP 2018), Brussels, Belgium. Association for Computational Linguistics.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019a). HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019b). Defending Against Neural Fake News. *arXiv: 1905.12616*.

Zhang, C., & Chai, J. (2009). What do We Know about Conversation Participants: Experiments on Conversation Entailment. In *Proceedings of the SIGDIAL 2009 Conference*, pp. 206–215. Association for Computational Linguistics.

Zhang, C., & Chai, J. (2010). Towards Conversation Entailment: An Empirical Investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 756–766. Association for Computational Linguistics.

Zhang, H., Liu, X., Pan, H., Song, Y., Wing-Ki, C., & Leung (2019). ASER: A Large-scale Eventuality Knowledge Graph. *arXiv: 1905.00270*.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., & Van Durme, B. (2018). ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv: 1810.12885*.

Zhang, S., Rudinger, R., Duh, K., & Van Durme, B. (2017). Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5, 379–395.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhou, B., Khashabi, D., Ning, Q., & Roth, D. (2019). "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, Hong Kong, China. Association for Computational Linguistics.

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7W: Grounded Question Answering in Images. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA. IEEE.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *IEEE International Conference on Compter Vision (ICCV 2015)*, pp. 19–27, Santiago, Chile. IEEE.