QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension

ANNA ROGERS, University of Copenhagen (Denmark), RIKEN (Japan)
MATT GARDNER, Allen Institute for Artificial Intelligence, USA
ISABELLE AUGENSTEIN, University of Copenhagen, Denmark

Alongside huge volumes of research on deep learning models in NLP in the recent years, there has been also much work on benchmark datasets needed to track modeling progress. Question answering and reading comprehension have been particularly prolific in this regard, with over 80 new datasets appearing in the past two years. This study is the largest survey of the field to date. We provide an overview of the various formats and domains of the current resources, highlighting the current lacunae for future work. We further discuss the current classifications of "reasoning types" in question answering and propose a new taxonomy. We also discuss the implications of over-focusing on English, and survey the current monolingual resources for other languages and multilingual resources. The study is aimed at both practitioners looking for pointers to the wealth of existing data, and at researchers working on new resources.

CCS Concepts: • **Computing methodologies** → **Language resources**; • **Information systems** → **Question answering**; *Information extraction.*

Additional Key Words and Phrases: reading comprehension, natural language understanding

ACM Reference Format:
Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. 1, 1 (July 2021), 38 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION: THE DATASET EXPLOSION

The rapid development of NLP data in the past years can be compared to the Cambrian explosion: the time when the fossil record shows a vast increase in the number of living species. In the case of NLP in 2013-2020, the key "resource" that made this explosion possible was the widespread use of crowdsourcing, essential for the new data-hungry deep learning models. The evolutionary forces behind the explosion were (a) a desire to push more away from linguistic structure prediction and towards a (still vague) notion of "natural language understanding" (NLU), which different research groups pursued in different directions, and (b) the increasing practical utility of commercial NLP systems incorporating questions answering technology (for search, chatbots, personal assistants, and other applications). A key factor in this process is that it was a breadth-first search: there was little coordination between groups (besides keeping track of concurrent work by competitor labs).

Authors' addresses: Anna Rogers, arogers@sodas.ku.dk, University of Copenhagen (Denmark), RIKEN (Japan); Matt Gardner, mattg@allenai.org, Allen Institute for Artificial Intelligence, USA; Isabelle Augenstein, augenstein@di.ku.dk, University of Copenhagen, Denmark.

Manuscript submitted to ACM

The result is a potpourri of datasets that is difficult to reduce to a single taxonomy, and for which it would be hard to come up with a single defining feature that would apply to all the resources. For instance, while we typically associate "question answering" (QA) and "reading comprehension" (RC) with a setting where there is an explicit question that the model is supposed to answer, even that is not necessarily the case. Some such datasets are in fact based on statements rather than questions (as in many cloze formatted datasets, see §3.2.3), or on a mixture of statements and questions.

The chief contribution of this work is a systematic review of the existing resources with respect to a set of criteria, which also broadly correspond to research questions NLP has focused on so far. After discussing the distinction between probing and information-seeking questions (§2), and the issue of question answering as a task vs format (§3.1), we outline the key dimensions for the format of the existing resources: questions (questions vs statements, §3.2), answers (extractive, multi-choice, categorical and freeform, §3.3), and input evidence (in terms of its modality, amount of information and conversational features, §3.4). Then we consider the domain coverage of the current QA/RC resources (§4) and the types of reasoning (§6), providing an overview of the current classifications and proposing our own taxonomy (along the dimensions of inference, information retrieval, world modeling, input interpretation, and multi-step reasoning). We conclude with the discussion of the issue of "requisite" skills and the gaps in the current research (§7).

For each of these criteria, we discuss how it is conceptualized in the field, with representative examples of English[1] resources of each type. What this set of criteria allows us to do is to place QA/RC work in the broader context of work on machine reasoning and linguistic features of NLP data, in a way that allows for easy connections to other approaches to NLU such as inference and entailment. It also allows us to map out the field in a way that highlights the cross-field connections (especially multi-modal NLP and commonsense reasoning) and gaps for future work to fill.

This survey focuses exclusively on the typology of the existing resources, and its length is proof that data work on RC/QA has reached the volume at which it is no longer possible to survey in conjunction with modeling work. We refer the reader to the existing surveys [204, 293] and tutorials [46, 223] for the current approaches to modeling in this area.

2 INFORMATION-SEEKING VS PROBING QUESTIONS

The most fundamental distinction in QA datasets is based on the communicative intent of the author of the question: was the person seeking information they did not have, or trying to test the knowledge of another person or machine?[2] Many questions appear "in the wild" as a result of humans seeking information, and some resources such as Natural Questions [137] specifically target such questions. Many other datasets consist of questions written by people who already knew the correct answer, for the purpose of probing NLP systems. These two kinds of questions broadly correlate with the "tasks" of QA and RC: QA is more often associated with information-seeking questions and RC with probing questions, and many of the other dimensions discussed in this survey tend to cluster based on this distinction.

Information-seeking questions tend to be written by users of some product, be it Google Search [53, 137], Reddit [79] or community question answering sites like StackOverflow [e.g. 36] and Yahoo Answers [e.g. 104], though there are some exceptions where crowd workers were induced to write information-seeking questions [54, 64, 82]. Most often, these questions assume no given context (§3.4.2) and are almost never posed as multiple choice (§3.3). Industrial research tends to focus on this category of questions, as research progress directly translates to improved products. An appealing aspect of these kinds of questions is that they typically arise from real-world use cases, and so can be sampled to create

[1] Most QA/RC resources are currently in English, so the examples we cite are in English, unless specified otherwise. §5 discusses the languages represented in the current monolingual and multilingual resources, including the tendencies & incentives for their creation.

[2] There are many other possible communicative intents of questions in natural language, such as expressing surprise, emphasis, or sarcasm. These do not as yet have widely-used NLP datasets constructed around them, so we do not focus on them in this survey.

a "natural" distribution of questions that people ask – this is why a dataset created from queries issued to Google Search was called "Natural Questions" [137]. Care must be taken in saying that there exists a "natural distribution" over all questions, however; the distribution of Google Search queries is in no way representative of all questions a person typically asks in a day, and it is not clear that such a concept is even useful, as questions have a wide variety of communicative intents.

Probing questions, on the other hand, tend to be written either by exam writers (§4) or crowd workers [e.g. 71, 210, 216]. The questions are most often written with the intent to probe understanding of a specific context, such as a paragraph or an image (§3.4.2); if a person were presented this context and wanted to extract some information from it, they would just examine the context instead of posing a question to a system. One could argue that questions written for testing human reading comprehension constitute a natural distribution of probing questions, but this distribution is likely not ideal for testing machine reading comprehension (see §3.3.2), especially if a large training set is given which can be mined for subtle spurious patterns. Instead, researchers craft classes of questions that probe particular aspects of reading comprehension in machines, and typically employ crowd workers to write large collections of these questions.

These two classes of questions also tend to differ in the kinds of reasoning they require (§6). Information-seeking questions are often ill-specified, full of "ambiguity and presupposition" [137], and so real-world QA applications would arguably need to show that they can handle this kind of data. But while the presence of ambiguous questions or questions with presuppositions make such data more "natural", it also makes these questions hard to use as benchmarks [33]: nearly half of the Natural Questions dataset are estimated to be ambiguous [175]. Especially when collected from search queries, information-seeking questions also tend to involve less complex reasoning than is seen in some probing datasets, as users do not expect search engines to be able to handle complex questions and so they do not ask them. This is not to say that there are no complex questions in search-based data, but they are more rare, while probing datasets can be specifically constructed to target one piece of the long tail in a more "natural" distribution.

Lastly, while we distinguish between information-seeking and probing questions, the lines are often blurry. For example, the question *"Which program at Notre Dame offers a Master of Education degree?"* could be asked by a college applicant seeking information, but it also occurs in SQuAD, a probing dataset [210]. When paired with single documents that likely contain the answer to the question, information-seeking datasets become much more like probing datasets [54]. Some datasets intentionally combine elements of both, probing an initial context while at the same time eliciting information seeking questions that need additional context to be answered [64, 82].

3 FORMAT

This section starts with the general discussion of when QA is a task and when it is a format (§3.1). Then, existing datasets are described along the dimension of formats for *questions* (the text used to query the system, §3.2), *answers* (the system output, §3.3), and *evidence* (the source of knowledge used to derive the system output, §3.4).

3.1 Task versus format

Strictly speaking, almost any NLP task can be formulated as question answering, and this is already being leveraged for model reuse and multi-task learning [e.g. 168, 271] and zero-shot learning [e.g. 1, 143]. For example, machine translation could be recast as answering questions like "What is the translation of X into German?", and sentiment analysis – as "What is the sentiment of X?". Under this view, a survey of QA datasets would encompass all NLP datasets.

The key distinction to keep in mind is "how easy would it be to replace the questions in a dataset with content-free identifiers?" [86]. An illustration of this heuristic is shown in Figure 1. Sentiment analysis is a classification task, so the

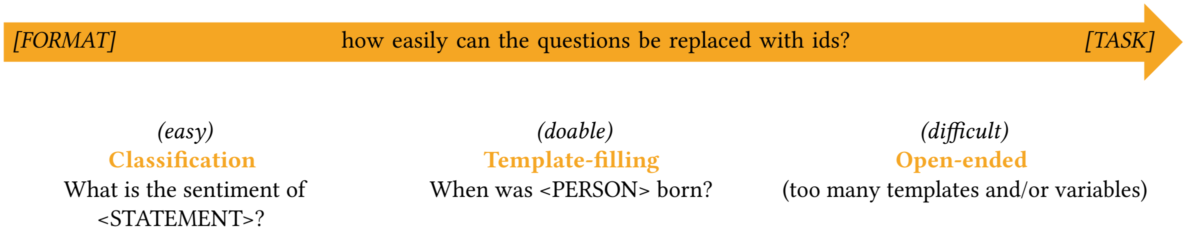
[FORMAT]
how easily can the questions be replaced with ids?
[TASK]
(easy)
Classification
What is the sentiment of <STATEMENT>?
(doable)
Template-filling
When was <PERSON> born?
(difficult)
Open-ended
(too many templates and/or variables)

Fig. 1. When is question answering a task, and when is it a format?

questions correspond to a few labels and could easily be replaced. An NLP system does not actually need to "understand" the wording of the recast question, beyond the part that needs to be classified. This heuristic is not a strict criterion, however, and the boundaries are fuzzy. Some datasets that have been published and used as QA or RC datasets can be templated with a few dozen templates [e.g. 272]. Still, such datasets have enabled undeniable progress, and will likely continue to be useful. What has changed is our awareness of how the low diversity of patterns in the training data leads to the over-reliance on these patterns [90, 118, 154, 171, among others].

One should also not conflate format with reasoning types (§6). For example, "extractive QA" is often discussed as if were a cohesive problem – however, extractive QA is an output format, and datasets using this format can differ wildly in the nature of the problems they encapsulate.

3.2 Question format

3.2.1 Natural language questions. Most QA and RC datasets have "questions" formulated as questions that a human speaker *could* ask, for either information-seeking or probing purposes (see §2). They could further be described in terms of their syntactic structure: yes/no questions (*Did it rain on Monday?*) wh-questions (*When did it rain?*), tag questions (It rained, didn't it?), or declarative questions (It rained?) [108]. Resources with syntactically well-formed questions as question format may come with any type of answer format described in §3.3.

3.2.2 Queries. Bordering on what could be characterized as "questions" linguistically are *queries*: the pieces of information that could be interpreted as a question (e.g. *tallest mountain Scotland* ⟶ *which mountain is the tallest in Scotland*?). Some QA resources (especially those with tables and knowledge bases (KB) as input) start with logically well-formed queries. From there, syntactically well-formed questions can be generated with templates [e.g. 272], and then optionally later edited by people [e.g. 96]. On the messy side of that spectrum we have search engine queries that people do *not* necessarily form as either syntactically well-formed questions or as KB queries. The current datasets with "natural" questions use filters to remove such queries [16, 137]. How we could study the full range of human interactions with search engines is an open problem at the boundary of QA and IR.

3.2.3 Cloze format. Cloze-format resources are based on statements rather than questions or queries: it is simply a sentence(s) with a masked span that, like in extractive QA (see §3.3.1), the model needs to predict. The key difference is that this statement is simply an excerpt from the evidence document (or some other related text), rather than something specifically formulated for information extraction. The sentences to be converted to Cloze "questions" have been identified as:

- simply sentences contained within the text [159];

Table 1. Answer formats of question answering and reading comprehension datasets

Evidence	Format	Question	Answer(s)	Example datasets
Einstein was born in 1879.	Extractive	When was Einstein born?	1879 (token 5)	SQuAD [210], NewsQA [256]
	Multi-choice	When was Einstein born?	(a) 1879, (b) 1880	RACE [138]
	Categorical	Was Einstein born in 1880?	No	BoolQ [53]
	Freeform	When was Einstein born?	1879 (generated)	MS MARCO [16], CoQA [213]

- designating an excerpt as the “text", and the sentence following it as the “question” [112, 194];
- given a text and summary of that text, use the summary as the question [110];

The Cloze format has been often used to test the knowledge of entities (CNN/Daily Mail [110], WikiLinks Rare Entity [159]). Other datasets targeted a mixture of named entities, common nouns, verbs (CBT [112], LAMBADA [194]). While the early datasets focused on single words or entities to be masked, there are also resources masking sentences in the middle of the narrative [62, 132].

The Cloze format has the advantage that these datasets can be created programmatically, resulting in quick and inexpensive data collection (although it can also be expensive if additional filtering is done to ensure answerability and high question quality [194]). But Cloze questions are not technically “questions”, and so do not *directly* target the QA task. The additional limitation is that only the relations within a given narrow context can be targeted, and it is difficult to control the kind of information that is needed to fill in the Cloze: it could simply be a collocation, or a generally-known fact – or some unique relation only expressed within this context.

The Cloze format is currently resurging in popularity also as a way to evaluate masked language models [78, 92], as fundamentally the Cloze task is what these models are doing in pre-training.

3.2.4 Story completion. A popular format in commonsense reasoning is the choice of the alternative endings for the passage (typically combined with multi-choice answer format (see §3.3.2)). It could be viewed as a variation of Cloze format, but many Cloze resources have been generated automatically from existing texts, while choice-of-ending resources tend to be crowdsourced for this specific purpose. Similarly to the Cloze format, the “questions” are not necessarily linguistically well-formed questions. They may be unfinished sentences (as in SWAG [284] and HellaSWAG [285]) or short texts (as in RocStories [181]) to be completed.

3.3 Answer format

The outputs of the current text-based datasets can be categorized as extractive (§3.3.1), multi-choice (§3.3.2), categorical (§3.3.3), or freeform (§3.3.4), as shown in Table 1.

3.3.1 Extractive format. Given a source of evidence and a question, the task is to predict the part of the evidence (a span, in case of a text) which is a valid answer for the question. This format is very popular both thanks to its clear relevance for QA applications, and the relative ease of creating such data (questions need to be written or collected, but answers only need to be selected in the evidence).

In its classic formulation extractive QA is the task behind search engines. The connection is very clear in early QA research: the stated goal of the first TREC QA competition in 2000 was “to foster research that would move retrieval systems closer to *information* retrieval as opposed to *document* retrieval” [265]. To answer questions like “Where is the Taj Mahal?” given a large collection of documents, the participating systems had to rank the provided documents and the candidate answer spans within the documents, and return the best five. Some of the more recent QA datasets also provide a collection of candidate documents rather than a single text (see §3.4.2).

A step back into this direction came with the introduction of unanswerable questions [3, 209]: the questions that target the same context as the regular questions, but do not have an answer in that context. With the addition of unanswerable questions, systems trained on extractive datasets can be used as a component of a search engine: first the candidate documents are assessed for whether they can be used to answer a given question, and then the span prediction is conducted on the most promising candidates. It is however possible to achieve search-engine-like behavior even without unanswerable questions [45, 52].

Many extractive datasets are "probing" in that the questions were written by the people who already knew the answer, but, as the datasets based on search engine queries show, it does not have to be that way. A middle ground is questions written by humans to test *human* knowledge, such as Trivia questions [124]: in this case, the writer still knows the correct answer, but the question is *not* written while seeing the text containing the answer, and so is less likely to contain trivial lexical matches.

The advantage of the extractive format is that only the questions need to be written, and the limited range of answer options means that it is easier to define what an acceptable correct answer is. The key disadvantage is that it limits the kinds of questions that can be asked to questions with answers directly contained in the text. While it is possible to pose rather complex questions (§6.2.5), it is hard to use this format for any interpretation of the facts of the text, any meta-analysis of the text or its author's intentions, or inference to unstated propositions.

3.3.2 Multi-choice format. Multiple choice questions are questions for which a small number of answer options are given as part of the question text itself. Many existing multi-choice datasets are expert-written, stemming from school examinations (e.g. RACE [138], CLEF QA [197]). This format has also been popular in datasets targeting world knowledge and commonsense information, typically based on crowdsourced narratives: MCTest [216], MCScript [192], RocStories [181].

The advantage of this format over the extractive one is that the answers are no longer restricted to something explicitly stated in the text, which enables a much wider range of questions (including commonsense and implicit information). The question writer also has full control over the available options, and therefore over the kinds of reasoning that the test subject would need to be capable of. This is why this format has a long history in human education. Evaluation is also straightforward, unlike with freeform answers. The disadvantage is that writing good multi-choice questions is not easy, and if the incorrect options are easy to rule out – the questions are not discriminative.[3]

Since multi-choice questions have been extensively used in education, there are many insights into how to write such questions in a way that would best test *human students*, both for low-level and high-level knowledge [15, 31, 163, 165]. However, it is increasingly clear that humans and machines do not necessarily find the same things difficult, which complicates direct comparisons of their performance. In particular, teachers are instructed to ensure that all the answer options items are plausible, and given in the same form [31, p.4]. This design could make the questions easy for a model backed with collocation information from a language model. However, NLP systems can be distracted by shallow lexical matches [118] or nonsensical adversarial inputs [267], and be insensitive to at least some meaning-altering perturbations [225]. For humans, such options that would be easy to reject.

[3] The STARC annotation scheme [24] is a recent proposal for controlling the quality of multi-choice questions by requiring that there are four answers, one of which is correct, one is based on a misunderstanding of the text span with the evidence for the correct answer, one is based on a distractor span, and one is plausible but unsupported by the evidence. This would allow studying the reasoning strategies of the models, but more studies are needed to show that we can generate these different types of incorrect answers at sufficient scale and without introducing extra spurious patterns.

Humans may also act differently when primed with different types of prior questions and/or when they are tired, whereas most NLP systems do not change between runs. QuAIL [221] made the first attempt to combine questions based on the textual evidence, world knowledge, and unanswerable questions, finding that this combination is difficult in human evaluation: if exposed to all three levels of uncertainty, humans have trouble deciding between making an educated guess and marking the question as unanswerable, while models do not.

3.3.3 Categorical format. We describe as "categorical" any format where the answers come from a strictly pre-defined set of options. As long as the set is limited to a semantic type with a clear similarity function (e.g. dates, numbers), we can have the benefit of automated evaluation metrics without the limitations of the extractive format. One could view the "unanswerable" questions in the extractive format [209] as a categorical task, which is then followed by answering the questions that can be answered (§3.3.1).

Perhaps the most salient example of the categorical answer format is boolean questions, for which the most popular resource is currently BoolQ [53]. It was collected as "natural" information-seeking questions in Google search queries similarly to Natural Questions [137]. Other resources not focusing on boolean questions specifically may also include them (e.g. MS MARCO [16], bAbI [272], QuAC [50]).

Another kind of categorical output format is when the set of answers seen during training is used as the set of allowed answers at test time. This allows for simple prediction – final prediction is a classification problem – but is quite limiting in that no test question can have an unseen answer. Visual question answering datasets commonly follow this pattern (e.g. VQA [8], GQA [115], CLEVR [123]).

3.3.4 Freeform format. The most natural setting for human QA is to generate the answer independently rather than choose from the evidence or available alternatives. This format allows for asking any kinds of question, and any other format can be instantly converted to it by having the system generate rather than select the available "gold" answer.

The problem is that the "gold" answer is probably not the only correct one, which makes evaluation difficult. Most questions have many correct or acceptable answers, and they would need to be evaluated on at least two axes: linguistic fluency and factual correctness. Both of these are far from being solved. On the factual side, it is possible to get high ROUGE-L scores on ELI5 [79] with answers conditioned on irrelevant documents [135], and even human experts find it hard to formulate questions so as to exactly specify the desired level of answer granularity, and to avoid presuppositions and ambiguity [33]. On the linguistic side, evaluating generated language is a huge research problem in itself [39, 261], and annotators struggle with longer answers [135]. There are also sociolinguistic considerations: humans answer the same question differently depending on the context and their background, which should not be ignored [219]).

So far the freeform format has not been very popular. Perhaps the best-known example is MS MARCO [16], based on search engine queries with human-generated answers (written as summaries of provided Web snippets), in some cases with several answers per query. Since 2016, the dataset has grown[4] to a million queries and is now accompanied with satellite IR tasks (ranking, keyword extraction). For NarrativeQA [131], crowd workers wrote both questions and answers based on book summaries. CoQA [213] is a collection of dialogues of questions and answers from crowd workers, with additional step for answer verification and collecting multiple answer variants. The writers were allowed to see the evidence, and so the questions are not information-seeking, but the workers were dynamically alerted to avoid words directly mentioned in the text. ELI5 [79] is a collection of user questions and long-form abstractive answers from the "Explain like I'm 5" subreddit, coupled with Web snippet evidence.

[4] https://microsoft.github.io/msmarco/

There is a lot of work to be done in the direction of of evaluation for freeform QA. As a starting point, Chen et al. [43] evaluate the existing automated evaluation metrics (BLEU, ROUGE, METEOR, F1) for extractive and multi-choice questions converted to freeform format, concluding that these metrics may be used for some of the existing data, but they limit the kinds of questions that can be posed, and, since they rely on lexical matches, they necessarily do poorly for the more abstractive answers. They argue for developing new metrics based on representation similarity rather than ngram matches [44], although the current implementations are far from perfect.

To conclude the discussion of answer formats in QA/RC, let us note that, as with other dimensions for characterizing existing resources, these formats do not form a strict taxonomy based on one coherent principle. Conceptually, the task of extractive QA could be viewed as a multi-choice one: the choices are simply all the possible spans in the evidence document (although most of them would not make sense to humans). The connection is obvious when these options are limited in some way: for example, the questions in CBT [112] are extractive (Cloze-style), but the system is provided with 10 possible entities from which to choose the correct answer, which makes also it a multi-choice dataset.

If the goal is general language understanding, we arguably do not even want to impose strict format boundaries. To this end, UnifiedQA [128] proposes a single "input" format to which they convert extractive, freeform, categorical (boolean) and multi-choice questions from 20 datasets, showing that cross-format training often outperforms models trained solely in-format.

3.4 Evidence format

By "evidence" or "context", we mean whatever the system is supposed to "understand" or use to derive the answer from (including but not limited to texts in natural language). QA/RC resources can be characterized in terms of the modality of their input evidence (§3.4.1), its amount (§3.4.2), and dynamic (conversational) vs static nature (§3.4.3).

3.4.1 Modality. While QA/RC is traditionally associated with natural language texts or structured knowledge bases, research has demonstrated the success of multi-modal approaches for QA (audio, images, and even video). Each of these areas is fast growing, and multimedia work may be key to overcoming issues with some implicit knowledge that is not "naturally" stated in text-based corpora [26].

Unstructured text. Most resources described as RC benchmarks [e.g. 210, 216] have textual evidence in natural language, while many QA resources come with multiple excerpts as knowledge sources (e.g. [16, 55]). See §3.4.2 for more discussions of the variation in the amount of text that is given as the context in a dataset.

Semi-structured text (tables). A fast-growing area is QA based on information from tables. At least four resources are based on Wikipedia tabular data, including WikiTableQuestions [195] and TableQA [260]. Two of them have supporting annotations for attention supervision: SQL queries in WikiSQL [290], operand information in WikiOps [49].

Structured knowledge. Open-domain QA with a structured knowledge source is an alternative approach to looking for answers in text corpora, except that in this case, the model has to explicitly "interpret" the question by converting it to a query (e.g. by mapping the text to a triplet of entities and relation, as in WikiReading [111]). The questions can be composed based on the target structured information, as in SimpleQuestions [32] or Event-QA [57]. The process is reversed in FreebaseQA [121], which collects independently authored Trivia questions and filters them to identify the subset that can be answered with Freebase information. The datasets may target a specific knowledge base: a general one such as WikiData [111] or Freebase [22, 121], or one restricted to a specific application domain [109, 274].

Images. While much of this work is presented in the computer vision community, the task of multi-modal QA (combining visual and text-based information) is a challenge for both computer vision and NLP communities. The complexity of the verbal component is on a sliding scale: from simple object labeling, as in MS COCO [153] to complex compositional questions, as in GQA [115].

While the NLP community is debating the merits of the "natural" information-seeking vs probing questions and both types of data are prevalent, (see §2), for visual QA the situation is skewed towards the probing questions, since most of them are based on large image bases such as COCO, Flickr or ImageNet which do not come with any independently occurring text. Accordingly, the verbal part may be created by crowdworkers based on the provided images (e.g. [242]), or (more frequently) generated, e.g. AQUA [84], IQA [95]. In VQG-Apple [196] the crowd workers were provided with an image and asked to write questions one *might* ask a digital assistant about that image, but the paper does not provide analysis of how realistic the result is.

Audio. "Visual QA" means "answering questions *about* images. Similarly, there is a task for QA *about* audio clips. DAQA [80] is a dataset consisting of audio clips and questions about what sounds can be heard in the audio, and in what order. As with most VQA work, the questions are synthetic.

Interestingly, despite the boom of voice-controlled digital assistants that answer users' questions (such as Siri or Alexa), public data for purely audio-based question answering is so far a rarity: the companies developing such systems undoubtedly have a lot of customer data, but releasing portions of it would be both ethically challenging and not aligned with their business interests. The result is that in audio QA the QA part seems to be viewed as a separate, purely text-based component of a pipeline with speech-to-text input and text-to-speech output. That may not be ideal, because in real conversations, humans take into account prosodic cues for disambiguation, but so far, there are few such datasets, making this a promising future research area. So far there are two small-scale datasets produced by human speakers: one based on TOEFL listening comprehension data [258], and one for a Chinese SquAD-like dataset [139]. Spoken-SQuAD [145] and Spoken-CoQA [282] have audio clips generated with a text-to-speech engine.

Another challenge for audio-based QA is the conversational aspect: questions may be formulated differently depending on previous dialogue. See §3.4.3 for an overview of the text-based work in that area.

Video. QA on videos is also a growing research area. Existing datasets are based on movies (MovieQA [251], MovieFIB [164]), TV shows (TVQA [141]), games (MarioQA [183]), cartoons (PororoQA [129]), and tutorials (TurorialVQA [56]). Some are "multi-domain": VideoQA [294] comprises clips from movies, YouTube videos and cooking videos, while TGIF-QA is based on miscellaneous GIFs [117].

As with other multimedia datasets, the questions in video QA datasets are most often generated [e.g. 117, 294] and the source of text used for generating those question matters a lot: the audio descriptions tend to focus on visual features, and text summaries focus on the plot [141]. TVQA questions are written by crowd workers, but they are still clearly probing rather than information-seeking. It is an open problem what a "natural" video QA would even be like: questions asked by someone who is deciding whether to watch a video? Questions asked to replace watching a video? Questions asked by movie critics?

Other combinations. While most current datasets fall into one of the above groups, there are also other combinations. For instance, HybridQA [47] and TAT-QA [292] target the information combined from text and tables, and MultiModalQA [250] adds images to that setting. MovieQA [251] has different "settings" based on what combination of input data is used (plots, subtitles, video clips, scripts, and DVS transcription).

The biggest challenge for all multimodal QA work is to ensure that all the input modalities are actually necessary to answer the question [253]: it may be possible to pick the most likely answer based only on linguistic features, or detect

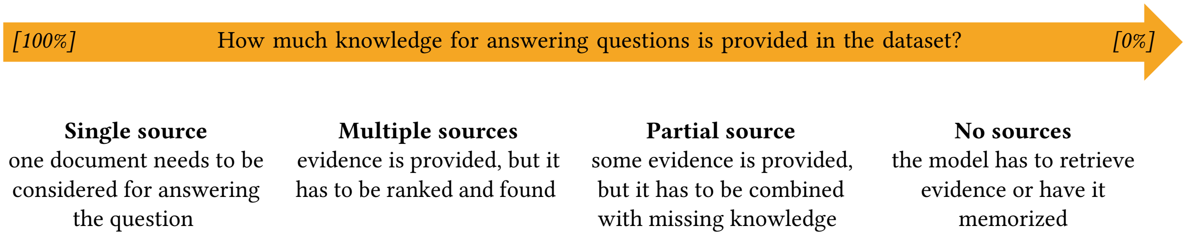
[100%]
How much knowledge for answering questions is provided in the dataset?
[0%]
Single source
one document needs to be considered for answering the question
Multiple sources
evidence is provided, but it has to be ranked and found
Partial source
some evidence is provided, but it has to be combined with missing knowledge
No sources
the model has to retrieve evidence or have it memorized

Fig. 2. Sources of knowledge for answering the questions.

the most salient object in an image while ignoring the question. After that, there is the problem of ensuring that *all* that multimodal information needs to be taken into account: for instance, if a model learns to answer questions about presence of objects based on a single image frame instead of the full video, it may answer questions incorrectly when the object is added/removed during the video. See also §7.1 for discussion of the problem of "required" skills.

3.4.2 Amount of evidence. The second dimension for characterizing the input of a QA/RC dataset is how much evidence the system is provided with. Here, we observe the following options:

- **Single source:** the model needs to consider a pre-defined tuple of a document and a question (and, depending on the format, answer option(s)). Most RC datasets such as RACE [138] and SQuAD [210] fall in this category. A version of this are resources with a *long* input text, such as complete books [131] or academic papers [64].
- **Multiple sources:** the model needs to consider a collection of documents to determine which one is the best candidate to contain the correct answer (if any). Many open-domain QA resources fall in this category: e.g. MS MARCO [16] and TriviaQA [124] come with retrieved Web snippets as the "texts". Similarly, some VQA datasets have multiple images as contexts [243].
- **Partial source:** The dataset provides documents that are necessary, but not sufficient to produce the correct answer. This may happen when the evidence snippets may be collected independently and not guaranteed to contain the answer, as in ARC [55]. Another frequent case is commonsense reasoning datasets such as RocStories [181] or CosmosQA [114]: there is a text, and the correct answer depends on both the information in this text and implicit world knowledge. E.g for SemEval2018 Task 11 [192] the organizers provided a commonsense reasoning dataset, and participants were free to use any external world knowledge resource.
- **No sources.** The model needs to rely only on some external source of knowledge, such as the knowledge stored in the weights of a pre-trained language model, a knowledge base, or an information retrieval component. A notable example is commonsense reasoning datasets, such as Winograd Schema Challenge [142] or COPA [94, 217]).

As shown in Figure 2, this is also more of continuum than a strict taxonomy. As we go from a single well-matched source of knowledge to a large heterogeneous collection, the QA problem increasingly incorporates an element of information retrieval. The same could be said for long single sources, such as long documents or videos, if answers can be found in a single relevant excerpt and do not require a high-level summary of the whole context. So far, our QA/RC resources tend to target more complex reasoning for shorter texts, as it is more difficult to create difficult questions over larger contexts.

Arguably, an intermediate case between single-source and multiple-source cases are datasets that collect multiple sources per question, but provide them already coupled with questions, which turns each example into a single-source problem. For example, TriviaQA [124] contains 95K questions, but 650K question-answer-evidence triplets.

3.4.3 Conversational features. The vast majority of questions in datasets discussed so far were collected or created as standalone questions, targeting a *static* source of evidence (text, knowledge base and/or any multimedia). The pragmatic context modeled in this setting is simply a set of standalone questions that could be asked in any order. But due to the active development of digital assistants, there is also active research on QA in conversational contexts: in addition to any sources of knowledge being discussed or used by the interlocutors, there is conversation history, which may be required to even interpret the question. For example, the question "Where did Einstein die"? may turn into "Where did he die?" if it is a follow-up question; after that, the order of the questions can no longer be swapped. The key differences to the traditional RC setting is that (a) the conversation history grows dynamically as the conversation goes on, (b) it is *not* the main source of information (that comes from some other context, a knowledge base, etc.).

While "conversational QA" may be intuitively associated with spoken (as opposed to written) language, the current resources for conversational QA do not necessarily originate in this way. For example, similarly to RC datasets like SQuAD, CoQA [213] was created in the written form, by crowd workers provided with prompts. It could be argued that the "natural" search engine queries have some spoken language features, but they also have their own peculiarities stemming from the fact that functionally, they are queries rather than questions (see §3.2).

The greatest challenge in creating conversational datasets is making sure that the questions are really information-seeking rather than probing (§2), since humans would not normally use the latter with each other (except perhaps in language learning contexts or checking whether someone slept through a meeting). From the perspective of how much knowledge the questioner has, existing datasets can be grouped into three categories:

- **Equal knowledge.** For example, CoQA [213] collected dialogues about the information in a passage (from seven domains) from two crowd workers, both of whom see the target passage. The interface discouraged the workers from using words occurring in the text.
- **Unequal knowledge.** For example, QuAC [50] is a collection of factual questions about a topic,[5] asked by one crowdworker and answered by another (who has access to a Wikipedia article). A similar setup to QuAC was used for the Wizards of Wikipedia [67], which, however, focuses on chitchat about Wikipedia topics rather than question answering, and could perhaps be seen as complementary to QuAC. ShARC [227] uses more than two annotators for authoring the main and follow-up questions to simulate different stages of a dialogue.
- **Repurposing "natural" dialogue-like data.** An example of this approach is Molweni [146], based on data from the Ubuntu Chat corpus, and its unique contribution is discourse level annotations in sixteen types of relations (comments, clarification questions, elaboration etc.) MANtIS [199] is similarly based on StackExchange dialogues, with a sample annotated for nine discourse categories. MSDialog [205] is based on Microsoft support forums, and the Ubuntu dialogue corpus [161] likewise contains many questions and answers from the Ubuntu ecosystem.

Again, these proposed distinctions are not clear-cut, and there are in-between cases. For instance, DoQA [36] is based on "real information needs" because the questions are based on StackExchange questions, but the actual questions

[5] In most conversational QA datasets collected in the unequal-knowledge setup the target information is factual, and the simulated scenario is that only one of the participants has access to that information (but theoretically, anyone could have such access). An interesting alternative direction is questions where the other participant is the only possible source of information: personal questions. CCPE-M [206] is a collection of dialogues where one party elicits the other party's movie preferences.

were still generated by crowdworkers in the “unequal knowledge” scenario, with real queries serving as “inspiration”. SHaRC [227] has a separate annotation step in which crowd workers formulate a scenario in which the dialogue they see could take place, i.e. trying to reverse-engineer the information need.

An emerging area in conversational QA is question rewriting:[6] rephrasing questions in a way that would make them easier to be answered e.g. through Web search results. CANARD [76] is a dataset of rewritten QuAC questions, and SaAC [7] is similarly based on a collection of TREC resources. QReCC [7] is a dataset of dialogues with seed questions from Natural Questions [137], and follow-up questions written by professional annotators. All questions come in two versions: the “natural” and search-engine-friendly version, e.g. by resolving pronouns to the nouns mentioned in the dialogue history. Disfl-QA [99] is a derivative of SQuAD with questions containing typical conversational “disfluencies" such as “uh” and self-corrections.

The above line of work is what one could call *conversational QA*. In parallel with that, there are datasets for *dialogue comprehension*, i.e. datasets for testing the ability to understand dialogues as opposed to static texts. They are “probing” in the same sense as e.g. RACE [138], the only difference being that the text is a dialogue script. In this category, FriendsQA [278] is based on transcripts of the ‘Friends’ TV show, with questions and extractive answers generated by crowd workers. There is also a Cloze-style dataset based on the same show [162], targeting named entities. DREAM [244] is a multi-choice dataset based on English exam data, with texts being dialogues.

Another related subfield is task-oriented (also known as goal-oriented) dialogue, which typically includes questions as well as transactional operations. The goal is for the user to collect information they need and then perform a certain action (e.g. find out what flights are available, choose and book one). There is some data for conversations with travel agents [116, 130], conducting meetings [6], navigation, scheduling and weather queries to an in-car personal assistant [77], and other [12, 232], as well as multi-domain resources [35, 200].

Conversational QA is actively studied not only in NLP, but also in information retrieval, and that community has produced many studies of actual human behavior in information-seeking dialogues that should be better known in NLP, so as to inform design of future resources. For instance, outside of maybe conference poster sessions and police interrogations, human dialogues do not usually consist only of questions and answers, which is e.g. the CoQA setting. Studies of human-system interaction [e.g. 255] elaborate on the types of conversational moves performed by the users (such as informing, rejecting, promising etc.) and how they could be modeled. In conversational QA there are also potentially many more signals useful in evaluation than simply correctness: e.g. MISC [252] is a small-scale resource produced by in-house MS staff that includes not only transcripts, but also audio, video, affectual and physiological signals, as well as recordings of search and other computer use and post-task surveys on emotion, success, and effort.

4 DOMAINS

One major source of confusion in the domain adaptation literature is the very notion of “domain”, which is often used to mean the source of data rather than any coherent criterion such as topic, style, genre, or linguistic register [211]. In the current QA/RC literature it seems to be predominantly used in the senses of “topic” and “genre” (a type of text, with a certain structure, stylistic conventions, and area of use). For instance, one could talk about the domains of programming or health, but either of them could be the subject of forums, encyclopedia articles, etc. which are “genres” in the linguistic sense. The below classification is primarily based on the understanding of “domain” as “genre”, with caveats where applicable.

[6] See also the task of “decontextualization” that could be used as“answer rewriting” [51]: in QA/RC, this means altering the sentences containing the answer so that they could be easily interpreted without reading the full text, e.g. by resolving coreference chains and replacing pronouns with nouns.

Encyclopedia. Wikipedia is probably the most widely used source of knowledge for constructing QA/RC datasets [e.g. 71, 111, 210, 277]. The QA resources of this type, together with those based on knowledge bases and Web snippets, constitute what is in some communities referred to as "open-domain" QA.[7] Note that here the term "domain" is used in the "topic" sense: Wikipedia, as well as Web and knowledge bases, contain much specialist knowledge, and the difference from the resources described below as "expert materials" is only that it is not restricted to particular topics.

Fiction. While fiction is one of the areas where large amounts of public-domain data is available, surprisingly few attempts were made to use them as reading comprehension resources, perhaps due to the incentive for more "useful" information-seeking QA work. CBT [112] is an early and influential Cloze dataset based on children's stories. BookTest [17] expands the same methodology to a larger number of project Gutenberg books. Being Cloze datasets, they inherit the limitations of the format discussed in §3.2.3. The first attempt to address a key challenge of fiction (understanding a long text) is NarrativeQA [131]: for this dataset, a QA system is supposed to answer the questions while taking into account the full text of a book. However, a key limitation of this data is that the questions were formulated based on book summaries, and thus are likely to only target major plot details.

The above resources target literary or genre fiction: long, complex narratives created for human entertainment or instruction. NLP papers also often rely on fictional mini-narratives written by crowdworkers for the purpose of RC tests. Examples of this genre include MCTest [216], MCScript [179, 191, 192], and RocStories [181].

Academic tests. This is the only "genre" outside of NLP where experts devise high-quality discriminative probing questions. Most of the current datasets were sourced from materials written by expert teachers to test students, which in addition to different subjects yields the "natural" division by student level (different school grades, college etc.). Arguably, it corresponds to level of difficulty of target concepts (if not necessarily language). Among the college exam resources, CLEF competitions [197, 198] and NTCIR QA Lab [235] were based on small-scale data from Japanese university entrance exams. RACE-C [149] draws on similar data developed for Chinese university admissions. ReClor [283] is a collection of reading comprehension questions from standartized admission tests like GMAT and LSAT, selected specifically to target logical reasoning.

In the school-level tests, the most widely-used datasets are RACE [138] and DREAM [244], both comprised of tests created by teachers for testing the reading comprehension of English by Chinese students (on narratives and multi-party dialogue transcripts, respectively). ARC [55] targets science questions authored for US school tests. OpenBookQA [174] also targets elementary science knowledge, but the questions were written by crowdworkers. ProcessBank [23] is a small-scale multi-choice dataset based on biology textbooks.

News. Given the increasing problem of online misinformation (see §7.3), question answering for news is a highly societally important area of research, but it is hampered by the lack of public-domain data. The best-known reading comprehension dataset based on news is undoubtedly the CNN/Daily Mail Cloze dataset [110], focusing on the understanding of named entities and coreference relations within a text. Subsequently NewsQA [256] also relied on CNN data; it is an extractive dataset with questions written by crowd workers. Most recently, NLQuAD [238] is an extractive benchmark with "non-factoid" questions (originally BBC news article subheadings) that need to be matched with longer spans within the articles. In multi-choice format, a section of QuAIL [221] is based on CC-licensed news. There is also a small test dataset of temporal questions for news events over a New York Times archive [269]

E-commerce. There are two e-commerce QA datasets based on Amazon review data. The earlier one was based on a Web crawl of questions and answers about products posed by users [167], and the more recent one (AmazonQA [101])

[7] In other communities, "open-domain" somewhat confusingly implies not something about a "domain" per se, but a format: that no evidence is given for a question, and that information must be retrieved from some corpus, which is often Wikipedia.

built upon it by cleaning up the data, and providing review snippets and (automatic) answerability annotation. SubjQA [28] is based on reviews from more sources than just Amazon, has manual answerability annotation and, importantly, is the first QA dataset to also include labels for subjectivity of answers.

Expert materials. This is a loose group of QA resources defined not by genre (the knowledge source is presumably materials like manuals, reports, scientific papers etc.), but by the narrow, specific topic only known to experts on that topic. This might be the most common category for QA datasets, since domain-specific chatbots for answering frequent user questions are increasingly used by companies, but these datasets are rarely made available to the research community.

Most existing resources are based on *answers provided by volunteer experts*: e.g. TechQA [38] is based on naturally-occurring questions from tech forums. A less common option is to hire experts, as done for Qasper [64]: a dataset of expert-written questions over NLP papers.

The "volunteer expert" setting is the focus of the subfield of *community QA*. It deserves a separate survey, but the key difference to the "professional" support resources is that the answers are provided by volunteers with varying levels of expertise, on platforms such as WikiAnswers [2], Reddit [79], or AskUbuntu [68]. Since the quality and amount of both questions and answers vary a lot, in that setting new QA subtasks emerge, including duplicate question detection and ranking multiple answers for the same question [184–186].

The one expert area with abundant expert-curated QA/RC resources is *biomedical QA*. BioASQ is a small-scale biomedical corpus targeting different NLP system capabilities (boolean questions, concept retrieval, text retrieval), that were initially formulated by experts as a part of CLEF competitions [197, 198, 257]. PubMedQA [122] is a corpus of biomedical literature abstracts that treats titles of articles as pseudo-questions, most of the abstract as context, and the final sentence of the abstract as the answer (with a small manually labeled section and larger unlabeled/artificially labeled section). In the healthcare area, CliCR [245] is a Cloze-style dataset of clinical records, and Head-QA [264] is a multimodal multi-choice dataset written to test human experts in medicine, chemistry, pharmacology, psychology, biology, and nursing. emrQA [193] is an extractive dataset of clinical records with questions generated from templates, repurposing annotations from other NLP tasks such as NER. There is also data specifically on the COVID pandemic [180].

Social media. Social media data present a unique set of challenges: the user speech is less formal, more likely to contain typos and misspellings, and more likely to contain platform-specific phenomena such as hashtags and usernames. So far there are not so many such resources. The most notable dataset in this sphere is currently TweetQA [275], which crowdsourced questions and answers for (news-worthy) tweet texts.

Multi-domain. Robustness across domains is a major issue in NLP, and especially in question answering where the models trained on one dataset do not necessarily transfer well to another even when within one domain [280]. However, so far there are very few attempts to create multi-domain datasets that could encourage generalization by design, and, as discussed above, they are not necessarily based on the same notion of "domain". In the sense of "genre", the first one was CoQA [213], combining prompts from children's stories, fiction, high school English exams, news articles, Wikipedia, science and Reddit articles. It was followed by QuAIL [221], a multi-choice dataset balanced across news, fiction, user stories and blogs.

In the sense of "topic", two more datasets are presented as "multi-domain": MMQA [100] is an English-Hindi dataset of Web articles that is presented as a multi-domain dataset, but is based on Web articles on the topics of tourism, history, diseases, geography, economics, and environment. In the same vein, MANtIS [199] is a collection of information-seeking dialogues from StackExchange fora across 14 topics (Apple, AskUbuntu, DBA, DIY, ELectronics, English, Gaming, GIS, Physics, Scifi, Security, Stats, Travel, World-building).

There are also “collective” datasets, formed as a collection of existing datasets, which may count as “multi-domain” by different criteria. In the sense of “genre”, ORB [70] includes data based on news, Wikipedia, fiction. MultiReQA [97] comprises 8 datasets, targeting textbooks, Web snippets, Wikipedia, scientific articles.

5 LANGUAGES

5.1 Monolingual resources

As in other areas of NLP, the “default” language of QA and RC is **English** [20], and most of this survey discusses English resources. The second best-resourced language in terms or QA/RC data is **Chinese**, which has the counterparts of many popular English resources. Besides SQuAD-like resources [59, 233], there is shared task data for open-domain QA based on structured and text data [72]. WebQA is an open-domain dataset of community questions with entities as answers, and web snippets annotated for whether they provide the correct answer [147]. ReCO [268] targets boolean questions from user search engine queries. There are also cloze-style datasets based on news, fairy tales, and children’s reading material, mirroring CNN/Daily Mail and CBT [60, 61], as well as a recent sentence-level cloze resource [62]. DuReader [107] is a freeform QA resource based on search engine queries and community QA. In terms of niche topics, there are Chinese datasets focusing on history textbooks [288] and maternity forums [276].

In the third place we have **Russian**, which a version of SQuAD [75], a dataset for open-domain QA over Wikidata [134], a boolean QA dataset [91], and datasets for cloze-style commonsense reasoning and multi-choice, multi-hop RC [81].

The fourth best resourced language is **Japanese**, with a Cloze RC dataset [270], a manual translation of a part of SQuAD [10], and a commonsense reasoning resource [189].

Three more languages have their versions of SQuAD [210]: **French** [66, 126], **Vietnamese** [187], and **Korean** [150], and there are three more small-scale evaluation sets (independently collected for **Arabic** [182]), human-translated to **French** [10]). **Polish** has a small dataset of open-domain questions based on Wikipedia “Did you know...?” data [166]. And, to the best of our knowledge, this is it: not even the relatively well-resourced languages like German necessarily have any monolingual QA/RC data. There is more data for individual languages that is part of multilingual benchmarks, but that comes with a different set of issues (§5.2).

In the absence of data, the researchers resort to machine translation of English resources. For instance, there is such SQuAD data for **Spanish** [37], **Arabic** [182], **Italian** [58], **Korean** [140]. However, this has clear limitations: machine translation comes with its own problems and artifacts, and in terms of content even the best translations could differ from the questions that would be “naturally” asked by the speakers of different languages.

The fact that so few languages have many high-quality QA/RC resources reflecting the idiosyncrasies and information needs of the speakers of their languages says a lot about the current distribution of funding for data development, and the NLP community appetite for publishing non-English data at top NLP conferences. There are reports of reviewer bias [220]: such work may be perceived as “niche” and low-impact, which makes it look like a natural candidate for second-tier venues[8], which makes such work hard to pursue for early career researchers.

This situation is not only problematic in terms of inclusivity and diversity (where it contributes to unequal access to the latest technologies around the globe). The focus on English is also counter-productive because it creates the wrong impression of progress on QA/RC vs the *subset* of QA/RC that is easy in English. For instance, as pointed out by the authors of TydiQA [54], questions that can be solved by string matching are easy in English (a morphologically poor language), but can be very difficult in languages with many morphophonological alternations and compounding.

[8] e.g. *Findings of EMNLP* was specifically created as a venue for which “there is no requirement for high perceived impact, and accordingly solid work in untrendy areas and other more niche works will be eligible” (https://2020.emnlp.org/blog/2020-04-19-findings-of-emnlp)

Another factor contributing to the perception of non-English work as "niche" and low-impact is that many such resources are "replications" of successful English resources, which makes them look derivative (see e.g. the above-mentioned versions of SQuAD). However, conceptually the contribution of such work is arguably comparable to incremental modifications of popular NLP architectures (a genre that does not seem to raise objections of low novelty), while having potentially much larger real-world impact. Furthermore, such work may also require non-trivial adaptations to transfer an existing methodology to a different language, and/or propose first-time innovations. For instance, MATINF [276] is a Chinese dataset jointly labeled for classification, QA and summarization, so that the same data could be used to train for all three tasks. The contribution of Watarai and Tsuchiya [270] is not merely a Japanese version of CBT, but also a methodology to overcome some of its limitations.

5.2 Multilingual resources

One way in which non-English work seems to be easier to publish is multilingual resources. Some of them are data from cross-lingual shared tasks[9], and also independent academic resources (such as English-Chinese cloze-style XCMRC [157]). But in terms of number of languages, the spotlight is currently on the following larger-scale resources:

- MLQA [144] targets extractive QA over Wikipedia with partially parallel texts in seven languages: English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. The questions are crowdsourced and translated.
- XQuAD [9] is a subset of SQuAD professionally translated into 10 languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi.
- XQA [156] is an open-domain QA dataset targeting entities; it provides training data for English, and test and development data for English and eight other languages: French, German, Portuguese, Polish, Chinese, Russian, Ukrainian, and Tamil.
- TydiQA [54] is the first resource of "natural" factoid questions in ten typologically diverse languages in addition to English: Arabic, Bengali, Finnish, Japanese, Indonesian, Kiswahili, Korean, Russian, Telugu, and Thai.
- XOR QA [11] builds on Tidy QA data to pose the task of cross-lingual QA: answering questions, where the answer data is unavailable in the same language as the question. It is a subset of TidyQA with data in seven languages: Arabic, Bengali, Finnish, Japanese, Korean, Russian and Telugu, with English as the "pivot" language (professionally translated).
- XQuAD-R and MLQA-R [222] are based on the above-mentioned XQuAD and MLQA extractive QA resources, recast as multilingual information retrieval tasks.
- MKQA [160] is based on professional translations of a subset of Natural Questions [137], professionally translated into 26 languages, focusing on "translation invariant" questions.

While these resources are a very valuable contribution, in multilingual NLP they seem to be playing the role similar to the role that the large-scale language models play in development of NLP models: the small labs are effectively out of the competition [218]. In comparison with large multilingual leaderboards, monolingual resources are perceived as "niche", less of a valuable contribution, less deserving of the main track publications on which careers of early-stage researchers depend. But such scale is only feasible for industry-funded research: of all the above multilingual datasets, only the smallest one (XQA) was not produced in affiliation with either Google, Apple, or Facebook.

Furthermore, scale is not necessarily the best answer: focus on multilinguality necessarily requires missing a lot of nuance that is only possible for in-depth work on individual languages performed by experts in those languages. A

[9] See e.g. QALD-4.1 [197], IJCNLP-2017 Task 5 [98].

key issue in multilingual resources is collecting data that is homogeneous enough across languages to be considered a fair and representative cross-lingual benchmark. That objective is necessarily competing with the objective of getting a natural and representative sample of questions in each individual language. To prioritize the latter objective, we would need comparable corpora of naturally occurring multilingual data. This is what happened in XQA [156] (based on the “Did you know... ?” Wikipedia question data), but there is not much such data that is in public domain. Tidy QA [54] attempts to approximate “natural” questions by prompting speakers to formulate questions for the topics, on which they are shown the header excerpts of Wikipedia articles, but it is hard to tell to what degree this matches real information needs, or samples all the linguistic phenomena that are generally prominent in questions for this language and should be represented.

A popular solution that sacrifices representativeness of individual languages for cross-lingual homogeneity is using translation, as it was done in MLQA [144], xQuaD [9], and MKQA [160]. However, translationese has many issues. In addition to the high cost, even the best human translation is not necessarily similar to naturally occurring question data, since languages differ in what information is made explicit or implicit [54], and cultures also differ in what kinds of questions typically get asked.

A separate (but related) problem is that it is also not guaranteed that translated questions will have answers in the target language data. This issue lead XQuAD to translating both questions and texts, MLQA – to partial cross-lingual coverage, MKQA – to providing only questions and answers, without the evidence texts, and XOR QA [11] – to positing the task of cross-lingual QA.

One more issue in multilingual NLP that does not seem to have received much attention in QA/RC research is code-switching [237], even though it clearly has a high humanitarian value. For instance, in the US context better question answering with code-switched English/Spanish data could be highly useful in the civil service and education, supporting the learning of immigrant children and social integration of their parents. So far there are only a few small-scale resources for Hindi [18, 40, 102, 207], Telugu and Tamil [40].

6 REASONING SKILLS

6.1 Existing taxonomies

We discussed above how different QA and RC datasets may be based on different understandings of “format” (§3.1) and “domain” (§4), but by far the least agreed-upon criterion is “types of reasoning”. While nearly every paper presenting a RC or QA dataset also presents some exploratory data analysis of a small sample of their data, the categories they employ vary too much to enable direct comparisons between resources.

Before discussing this in more detail, let us recap how “reasoning” is defined. In philosophy and logic “any process of drawing a conclusion from a set of premises may be called a process of reasoning” [30]. Note that this is similar to the definition of “inference”: “the process of moving from (possibly provisional) acceptance of some propositions, to acceptance of others” [29]. But this definition does not cover everything that is discussed as “reasoning types” or “skills” in QA/RC literature.

To date, two comprehensive taxononomies for the QA/RC “skills” have been proposed in the NLP literature:

- Sugawara and Aizawa [239], Sugawara et al. [240] distinguish between object tracking skills, mathematical reasoning, logical reasoning, analogy, causal and spatiotemporal relations, ellipsis, bridging, elaboration, meta-knowledge, schematic clause relation, punctuation.

- Schlegel et al. [231] distinguish between operational (bridge, constraint, comparison, intersection), arithmetic (subtraction, addition, ordering, counting, other), and linguistic (negation, quantifiers, conditional monotonicity, con/dis-junction) meta-categories, as opposed to temporal, spatial, causal reasoning, reasoning "by exclusion" and "retrieval". They further describe questions in terms of knowledge (factual/intuitive) and linguistic complexity (lexical and syntactic variety, lexical and syntactic ambiguity).

A problem with any taxonomy is that using it to characterize new and existing resources involves expensive fine-grained expert annotation. A frequently used workaround is a kind of keyword analysis by the initial words in the question (since for English that would mean *what*, *where*, *when* and other question words). This was done e.g. in [e.g. 16, 131, 192], and Dzendzik et al. [74] perform such an analysis across 37 datasets, showing that 22% of all questions are "what" questions. However, it is a characterization of the *answers* to the questions, rather than *the process* used to answer the question. It is also a very crude heuristic for the semantic type: for instance, "what" questions could target not only entities, but also properties (*what color?*), locations (*what country?*), temporal information (*what day?*), etc.

6.2 Proposed taxonomy

Based on [231, 239, 240], we propose an alternative taxonomy of QA/RC "skills" along the following dimensions:

- **Inference** (§6.2.1): "the process of moving from (possibly provisional) acceptance of some propositions, to acceptance of others" [29].
- **Retrieval** (§6.2.2): knowing where to look for the relevant information.
- **Input interpretation & manipulation** (§6.2.3): correctly understanding the meaning of all the signs in the input, both linguistic and numeric, and performing any operations on them that are defined by the given language/mathematical system (identifying coreferents, summing up etc.).
- **World modeling** (§6.2.4): constructing a valid representation of the spatiotemporal and social aspects of the world described in the text, as well as positioning/interpreting the text itself with respect to the reader and other texts.
- **Multi-step** (§6.2.5): performing chains of actions on any of the above dimensions.

A key feature of the taxonomy is that these dimensions are *orthogonal*: the same question can be described in terms of their linguistic form, the kind of inference required to arrive at the answer, retrievability of the evidence, compositional complexity, and the level of world modeling (from generic open-domain questions to questions about character relations in specific books). In a given question, some of them may be more prominent/challenging than others.

Our proposal is shown in Figure 3 and discussed in more detail below.

6.2.1 Inference type. Fundamentally, the problem of question answering can be conceptualized as the classification of the relation between the premise (context+question) and the conclusion (a candidate answer) [226]. Then, the type of reasoning performed by the system can be categorized in the terms developed in logic and philosophy[10]. Among the criteria developed for describing different types of reasoning is the *direction of reasoning*: deductive (from premise to conclusion) and abductive (from conclusion to the premise that would best justify the conclusion) [69]. Another key criterion is *the degree to which the premise supports the conclusion*: in deductive reasoning, the hypothesis is strictly entailed by the premise, and in inductive reasoning, the support is weaker [106]. Finally, reasoning could be analysed

[10] Note that logical reasoning is only a subset of human reasoning: human decisions are not necessarily rational, they may be based on biases, heuristics, or fall pray to different logical fallacies. It is not clear to what extent the human reasoning "shortcuts" should be replicated in machine RC systems, especially given that they develop their own biases and heuristics (see §7.1)

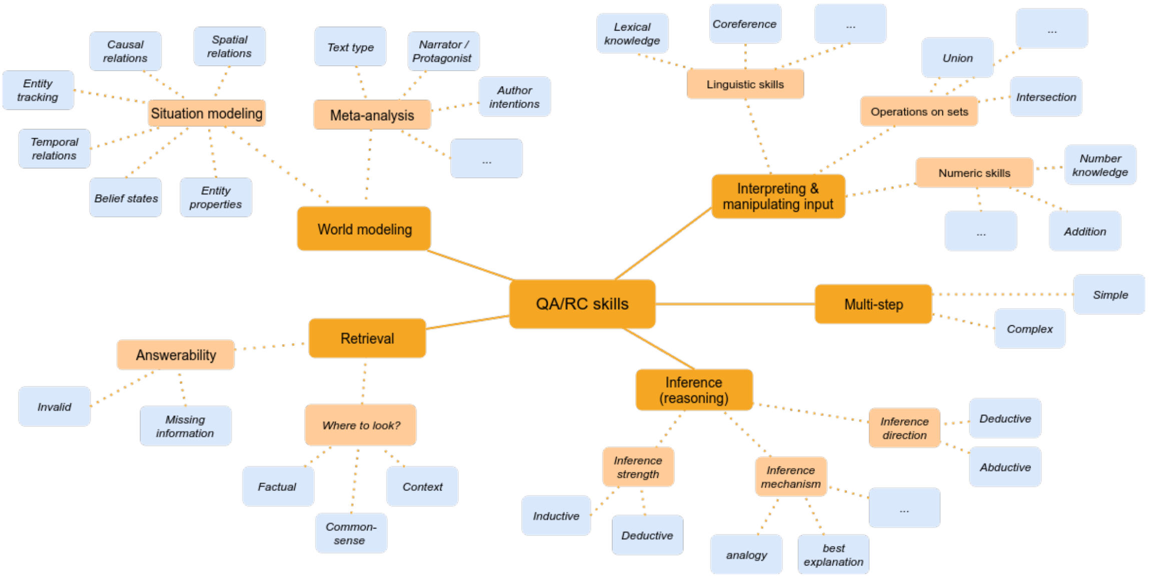

Fig. 3. Proposed classification of machine reading comprehension skills

with respect to the *kind of support for the conclusion*, including analogical reasoning [19], defeasible reasoning ("what normally[11] happens", [133]), and "best explanation" [69].

While the above criteria are among the most fundamental and well-recognized to describe human reasoning, none of them is actively used to study machine reasoning, at least in the current QA/RC literature. Even though deductive reasoning is both fundamental and the most clearly mappable to what we could expect from machine reasoning, to the best of our knowledge so far there is only one dataset for that: LogiQA [155], a collection of multi-choice questions from civil servant exam materials.

To further complicate the matter, sometimes the above-mentioned terms are even used differently. For instance, ReClor [283] is presented as a resource targeting logical reasoning, but it is based on GMAT/LSAT teaching materials, and much of it actually targets meta-analysis of the logical structure rather than logical reasoning itself (e.g. identifying claims and conclusions in the provided text). CLUTTR [236] is an inductive reasoning benchmark for kinship relations, but the term "inductive" is used in the sense of "inducing rules" (similar to the above definition of "inference") rather than as "non-deductive" (i.e. offering only partial support for the conclusion).

A kind of non-deductive reasoning that historically received a lot of attention in the AI literature is defeasible reasoning [48, 169], which is now making a comeback in NLI [224] (formulated as the task of re-evaluating the strength of the conclusion in the light of an additional premise strengthening/weakening the evidence offered by the original premise). There is also ART [25], an abductive reasoning challenge where the system needs to come up with a hypothesis that better complements incomplete observations.

6.2.2 Retrieval. It could be argued that information retrieval happens *before* inference: to evaluate a premise and a conclusion, we first have to have them. However, inference can also be viewed as the *ranking mechanism* of retrieval:

[11] There is debate on whether probabilistic inference belongs to inductive or deductive logic [65], which we leave to the philosophers. But SEP defines defeasible reasoning as non-deductive reasoning based on "what normally happens" [133], which seems to presuppose the notion of probability.

Example	*Problem*
Have you stopped beating your wife?	invalid premise (that the wife is beaten)
What is the meaning of life, the universe, and everything? [5]	not specific enough
At what age can I get a driving license?	missing information (in what country?)
Can quantum mechanics and relativity be linked together?	information not yet discovered
What was the cause of the US civil war? [33]	no consensus on the answer
Who can figure out the true meaning of ‘covfefe’?	uninterpretable due to language errors
Do colorless ideas sleep furiously?	syntactically well-formed but uninterpretable
What is the sum of angles in a triangle with sides 1, 1, and 10 cm?[14]	such a triangle cannot exist
What have the Romans ever done for us? [41]	rhetorical question
What is the airspeed velocity of a swallow carrying a coconut? [273]	the answer would not be useful[15]to know

Table 2. Types of invalid questions

NLP systems consider the answer options so as to choose the one offering the strongest support for the conclusion. This is how the current systems approach close-world reading comprehension tests like RACE [138] or SQuAD [210]. In the open-world setting, instead of a specific text, we have a much broader set of options (a corpus of snippets, a knowledge base, knowledge encoded by a language model etc.). However, fundamentally the task is still to find the best answer out of the available knowledge. We are considering two sub-dimensions of the retrieval problem: determining whether an answer exists, and where to look for it.

Answerability. SQuAD 2.0 [209] popularized the distinction between questions that are answerable with the given context and those that are not. However, the distinction is arguably not binary, and at least two resources argue for a 3-point uncertainty scale. ReCO [268] offers boolean questions with “yes”, “no” and “maybe” answer options. QuAIL [221] distinguishes between full certainty (answerable with a given context), partial certainty (a confident guess can be made with a given context + some external common knowledge), and full uncertainty (no confident guess can be made even with external common knowledge). A more general definition of the unanswerable questions would be this: the questions that cannot be answered *given all the information that the reader has access to.*

This is different from *invalid questions*: the questions that a human would reject rather than attempt to answer. Table 2 shows examples for different kinds of violations: the answers that are impossible to retrieve, loaded questions, ill-formed questions, rhetorical questions, “useless” questions, and others.

Where to look for the target knowledge? The classical RC case in resources like SQuAD is a single *context* that is the only possible source of information: in this case, the retrieval problem is reduced to finding the relevant span. When the knowledge is not provided, the system needs to know where to find it,[12] and in this case it may be useful to know whether it is *factual* (e.g. “Dante was born in Florence”) or *world knowledge* (e.g. “bananas are yellow”).[13] This is the core distinction between the subfields of open-domain QA and commonsense reasoning, respectively. Note that in both of those cases, the source of knowledge is external to the question and must be retrieved from somewhere (Web snippets, knowledge bases, model weights, etc.). The difference is in the *human* competence: an average human speaker is not expected to have all the factual knowledge, but is expected to have a store of the world knowledge (even though the specific subset of that knowledge is culture- and age-dependent).

[12] For human readers, McNamara and Magliano [172] similarly distinguish between *bridging* (linking new information—in this case, from the question—to previous context) and *elaboration* (linking information to some external information).

[13] Schlegel et al. [230] distinguish between “factual” and “intuitive” knowledge. The latter is defined as that “which is challenging to express as a set of facts, such as the knowledge that a parenthetic numerical expression next to a person’s name in a biography usually denotes [their] life span”.

Many resources for the former were discussed in §4. Commonsense reasoning resources deserve a separate survey, but overall, most levels of description discussed in this paper also apply to them. They have the analog of open-world factoid QA (e.g. CommonsenseQA [249], where the task is to answer a multi-choice question without any given context), but more resources are described as "reading comprehension", with multi-choice [114, 192] or cloze-style [286] questions asked in the context of some provided text. Similarly to "domains" in open-world QA (see §4), there are specialist resources targeting specific types of world knowledge (see §6.2.4).

6.2.3 Interpreting & manipulating input. This dimension necessarily applies to any question: both humans and machines *should* have the knowledge of the meaning of the individual constituent elements of the input (words, numbers), and have the ability to perform operations on them that are defined by the language/shared mathematical system (rather than given in the input).[16] It includes the following subcategories:

- **Linguistic skills.** SQuAD [210], one of the first major RC resources, predominantly targeted argument extraction and event paraphrase detection. Curently many resources focus on coreference resolution (e.g. Quoref [63], part of DROP [71]). Among the reasoning types proposed in [239, 240], "linguistic skills" also include ellipsis, schematic clause relations, punctuation. The list is not exhaustive: arguably, any questions formulated in a natural language depend on a large number of linguistic categories (e.g. reasoning about temporal relations must involve knowledge of verb tense), and even the questions targeting a single phenomenon as it is defined in linguistics (e.g. coreference resolution) do also require other linguistic skills (e.g. knowledge of parts of speech). Thus, any analysis based on linguistic skills should allow the same question to belong to several categories, and it is not clear whether we can reliably determine which of them are more "central".
  Questions (and answers/contexts) could also be characterized in terms of "ease of processing" [172], which is related to the set of linguistic phenomena involved in its surface form. But it probably does not mean the same thing for humans and machines: the latter have a larger vocabulary, do not get tired in the same way, etc.
- **Numeric skills.** In addition to the linguistic knowledge required for interpreting numeral expressions, an increasing number of datasets is testing NLP systems' abilities of answering questions that require mathematical operations over the information in the question and the input context. DROP [71] involves numerical reasoning over multiple paragraphs of Wikipedia texts. Mishra et al. [177] contribute a collection of small-scale numerical reasoning datasets including extractive, freeform, and multi-choice questions, some of them requiring retrieval of external world knowledge. There is also a number of resources targeting school algebra word problems [136, 173, 234, 259], and multimodal counting benchmarks [4, 42].
- **Operations on sets.** This category targets such operations as union, intersection, ordering, and determining subset/superset relations which going beyond the lexical knowledge subsumed by the hypernymy/hyponymy relations. The original bAbI [272] included "lists/sets" questions such as *Daniel picks up the football. Daniel drops the newspaper. Daniel picks up the milk. John took the apple. What is Daniel holding? (milk, football).* Among the categories proposed by Schlegel et al. [230], the "constraint" skill is fundamentally the ability to pick a subset the members which satisfy an extra criterion.

[12] https://philosophy.stackexchange.com/questions/37311/are-all-answers-to-a-contradictory-question-correct-or-are-all-wrong-or-is-it

[13] The practical utility of questions is hard to estimate objectively, given the wide range of human interests (especially cross-culturally). Horbach et al. [113] annotate questions for centrality to the given topic, and whether a teacher would be likely to use that question with human students, but the human agreement on their sample is fairly low. The agreement is likely even less for the more niche, specialist questions: the low agreement on acceptance recommendations in peer review [201] is likely partly due to the fact that different groups of researchers simply do not find each other's research questions equally exciting.

[16] The current NLP systems can perform well on QA/RC benchmarks even when they are transformed to become uninterpretable to humans [241]. It is an open question whether we should strive for systems to reject inputs that a human would reject, and on the same grounds.

Some linguistic phenomena highly correlate with certain reasoning operations, but overall these two dimensions are still orthogonal. A prime example is comparison:[17] it is often expressed with comparative degrees of adjectives (in the question or context) and so requires interpretation of those linguistic signs. At the same time, unless the answer is directly stated in the text, it also requires a deductive inference operation. For example: *John wears white, Mary wears black. Who wears darker clothes?*

6.2.4 World modeling. One of the key psychological theories of human RC is based on mental simulation: when we read, we create a model of the described world, which requires that we "instantiate" different objects and entities, track their locations, and ingest and infer the temporal and causal relations between events [262, 295]. Situation modeling has been proposed as one of the levels of representation in discourse comprehension [263], and it is the basis for the recent "templates of understanding" [73] that include spatial, temporal, causal and motivational elements. We further add the category of belief states [221], since human readers keep track not only of spatiotemporal and causal relations in a narrative, but also the who-knows-what information.

A challenge for psychological research is that different kinds of texts have a different mixture of prominent elements (temporal structure for narratives, referential elements in expository texts etc.), and the current competing models were developed on the basis of different kinds of evidence, which makes them hard to reconcile [172]. This is also the case for machine RC, and partly explains the lack of agreement about classification of "types of reasoning" across the literature. Based on our classification, the following resources explicitly target a specific aspect of situation modeling, in either RC (i.e. "all the necessary information in the text") or commonsense reasoning (i.e. "text needs to be combined with extra world knowledge") settings:[18]

- spatial reasoning: bAbI [272], SpartQA [176], many VQA datasets [e.g. 117, see §3.4.1]
- temporal reasoning: event order (QuAIL [221], TORQUE [188]), event attribution to time (TEQUILA [120], TempQuestions [119], script knowledge (MCScript [192]), event duration (MCTACO [291], QuAIL [221]), temporal commonsense knowledge (MCTACO [291], TIMEDIAL [203]), some multimodal datasets [80, 117]
- belief states: Event2Mind [212], QuAIL [221]
- causal relations: ROPES [152], QuAIL [221], QuaRTz [247].
- tracking entities: across locations (bAbI [272]), in coreference chains (Quoref [63], resources in the Winograd Schema Challenge family [142, 228]). Arguably the cloze-style resources based on named entities also fall into this category (CBT [112], CNN/DailyMail [110], WhoDidWhat [190]), but they do not guarantee that the masked entity is in some complex relation with its context.
- entity properties and relations:[19] social interactions (SocialIQa [229]), properties of characters (QuAIL [221]), physical properties (PIQA [27], QuaRel [246]), numerical properties (NumberSense [151]).

The text + alternative endings format used in several commonsense datasets like SWAG (see §3.2.4) has the implicit question "What happened next?". These resources cross-cut causality and temporality: much of such data seems to target causal relations (specifically, the knowledge of possible effects of interactions between characters and entities), but also script knowledge, and the format clearly presupposes the knowledge of the temporal before/after relation.

A separate aspect of world modeling is the meta-analysis skills: the ability of the reader to identify the likely time, place and intent of its writer, the narrator, the protagonist/antagonist, identifying stylistic features and other categories.

[17] So far comparison is directly targeted in QuaRel [246], and also present in parts of other resources [71, 231].

[18] This list includes only the resources that specifically target a given type of information, or have a part targeting that type of information that can be separated based on the provided annotation.

[19] QA models are even used directly for relation extraction [1, 143, 148].

These skills are considered as a separate category by Sugawara et al. [240], and are an important target of the field of literary studies, but so far they have not been systematically targeted in machine RC. That being said, some existing resources include questions formulated to include words like "author" and "narrator" [221]. They are also a part of some resources that were based on existing pedagogical resources, such as some of ReClor [283] questions that focus on identifying claims and conclusions in the provided text.

6.2.5 Multi-step reasoning. Answering a question may require one or several pieces of information. In the recent years a lot of attention was drawn to what could be called multi-step information retrieval, with resources focusing on "simple" and "complex" questions:

- "Simple" questions have been defined as such that "refer to a single fact of the KB" [32]. In an RC context, this corresponds to the setting where all the necessary evidence is contained in the same place in the text.
- The complex questions, accordingly, are the questions that rely on several facts [248]. In an RC setting, this corresponds to the so-called multi-hop datasets that necessitate the combination of information across sentences [127], paragraphs [71], and documents [279]. It also by definition includes questions that require a combination of context and world knowledge [e.g. 221].

That being said, the "multi-step" skill seems broader than simply combining several facts. Strictly speaking, any question is linguistically complex just because it is a compositional expression. We use some kind of semantic parsing to find the missing pieces of information for the question "Who played Sherlock Holmes, starred in Avengers and was born in London?" – but we must rely on the same mechanism to interpret the question in the first place. We may likewise need to perform several inference steps to retrieve the knowledge if it is not explicitly provided, and we regularly make chains of guesses about the state of the world (Sherlock Holmes stories exemplify a combination of these two dimensions).

7 DISCUSSION

This section concludes the paper with broader discussion of reasoning skills: the types of "skills" that are minimally required for our systems to solve QA/RC benchmarks (§7.1) vs the ones that a human would use (§7.2). We then proceed to highlighting the gaps in the current research, specifically the kinds of datasets that have not been made yet (§7.3).

7.1 What reasoning skills are actually required?

A key assumption in the current analyses of QA/RC data in terms of the reasoning skills they target (including our own taxonomy in §6.2) is that the skills that a *human* would use to answer a given question are also the skills that a *model* would use. However, that is not necessarily true. Fundamentally, DL models search for patterns in the training data, and they may and do find various shortcuts that happen to also predict the correct answer [90, 103, 118, 241, inter alia]. An individual question may well target e.g. coreference, but if it contains a word that is consistently associated with the first answer option in a multi-choice dataset, the model could potentially answer it without knowing anything about coreference. What is worse, *how* a given question is answered could change with a different split of the same dataset, a model with a different inductive bias, or, the most frustratingly, even a different run of the same model [170].

This means that there is a discrepancy between the reasoning skills that a question seems to target, and the skills that are minimally required to "solve" a particular dataset. In the context of the traditional machine learning workflow with training and testing, *we need to reconsider the idea that whether or not a given reasoning skill is "required" is a characteristic of a given question. It is rather a characteristic of the combination of that question and the entire dataset.*

The same limitation applies to the few-shot- or in-context-learning paradigm based on extra-large language models [34], where only a few samples of the target task are presented as examples and no gradient updates are performed. Conceptually, such models still encapsulate the patterns observed in the language model training data, and so may still be choosing the correct answer option e.g. because there were more training examples with the correct answer listed first (see [289] for an approach to countering this bias). The difference is only that it is much harder to perform the training data analysis and find any such superficial hints.

How can we ever tell that the model is producing the correct answer for the right reasons? There is now enough work in this area to deserve its own survey, but the main directions are roughly as follows:

- construction of *diagnostic tests*: adversarial tests [e.g. 118, 266], probes for specific linguistic or reasoning skills [e.g. 215, 236], minimal pair evaluation around the model's decision boundary [e.g. 85, 125].
- creating *larger collections of generalization tests*, both out-of-domain (§4) and cross-lingual (§5.2). The assumption is that as their number grows the likelihood of the model solving them all with benchmark-specific heuristics decreases.
- work on *controlling the signal in the training data*, on the assumption that if a deep learning model has a good opportunity to learn some phenomenon, it should do so (although that is not necessarily the case [89]). This direction includes all the work on resources focusing on specific reasoning or linguistic skills ([e.g. 63]) and balanced sets of skills [e.g. 221].

  This direction also includes the methodology work on crafting the data to avoid reasoning shortcuts: e.g. using human evaluation to discard the questions that humans could answer without considering full context [194].
- *interpretability work* on generating human-interpretable explanations for a given prediction, e.g. by context attribution [e.g. 214] or influential training examples [e.g. 208]. However, the faithfulness of such explanations is itself an active area of research [e.g. 202, 281]. The degree to which humans can use explanations to evaluate the quality of the model also varies depending on the model quality and prior belief bias [93].

While we may never be able to say conclusively that a blackbox model relies on the same strategies as a human reader, we should (and, under the article 13 of the AI Act proposal, could soon be legally required to[20]) at least identify the cases in which they succeed and in which they fail, as it is prerequisite for safe deployment.

7.2 Analysis of question types and reasoning skills

Section 7.1 discussed the fundamental difficulties with identifying how a blackbox neural model was able to solve a QA/RC task. However, we also have trouble even identifying the processes a *human* reader would use to answer a question. As discussed in §6.1, there are so far only two studies attempting cross-dataset analysis of reasoning skills according to a given skill taxonomy, and they both only target small samples (50-100 examples per resource). This is due to the fact that such analysis requires expensive expert annotation. Horbach et al. [113] showed that crowdworkers have consistently lower agreement even on annotating question grammaticality, centrality to topic, and the source of information for the answers. What is worse, neither experts nor crowdworkers were particularly successful with annotating "types of information needed to answer this question".

The dimension of our taxonomy (§6.2) that has received the least attention so far seems to be the logical aspects of the inference performed. Perhaps not coincidentally, this is the most abstract dimension requiring the most specialist knowledge. However, the logical criterion of the strength of support for the hypothesis is extremely useful: to be able to

[20] https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

trust NLP systems in the real world, we would like to know how they handle reasoning with imperfect information. This makes analysis of the "best-case" inference (assuming a non-heuristic reasoning process based on full interpretation of the questions and inputs) in the existing QA/RC resources a promising direction for future work. It could be bootstrapped by the fact that at least some of the inference types map to the question types familiar in QA/RC literature:

- The questions that involve only interpreting and manipulating (non-ambiguous) linguistic or numerical input fall under deductive reasoning, because the reader is assumed to have a set of extra premises (definitions for words and mathematical operations) shared with the question author.
- The questions about the future state of the world, commonsense questions necessarily have a weaker link between the premise and conclusion, and could be categorized as inductive.
- Other question types could target inductive or deductive reasoning, depending on how strong is the evidence provided in the premise: e.g. temporal questions are deductive if the event order strictly follows from the narrative, and inductive if there are uncertainties filled on the basis of script knowledge.

7.3 What datasets have not been created?

Notwithstanding all of the numerous datasets in the recent years, the space of unexplored possibilities remains large. Defining what datasets need to be created is itself a part of the progress towards machine NLU, and any such definitions will necessarily improve as we make such progress. At this point we would name the following salient directions.

Linguistic features of questions and/or the contexts that they target. The current pre-trained language models do not acquire all linguistic knowledge equally well or equally fast: e.g. RoBERTa [158] learns the English irregular verb forms already with 100M tokens of pre-training, but struggles with (generally more rare) syntactic island effects even after 1B pre-training [287]. Presumably knowledge that is less easily acquired in pre-training will also be less available to the model in fine-tuning. There are a few datasets that focus on questions requiring a specific aspect of linguistic reasoning, but there are many untapped dimensions. How well do our models cope with questions that a human would answer using their knowledge of e.g. scope resolution, quantifiers, or knowledge of verb aspect?

Pragmatic properties of questions. While deixis (contextual references to people, time and place) clearly plays an important role in multimodal and conversational QA resources, there does not seem to be much work focused on that specifically (although many resources cited in §3.4.3 contain such examples). Another extremely important direction is factuality: there is already much research on fact-checking [13, 14, 105, 254], but beyond that, it is also important to examine questions for presuppositions.

QA for the social good. A very important dimension for practical utility of QA/RC data is their domain (§4): domain adaptation is generally very far from being solved, and that includes the transfer between QA/RC datasets [83, 280]. There are many domains that have not received much attention because they are not backed by commercial interests, and are not explored by academics because there is no "wild" data like StackExchange questions that could back it up. For instance, QA data that could be used to train FAQ chatbots for education and nonprofit sectors could make a lot of difference for low-resource communities, but is currently notably absent. And beyond the English-speaking world, high-quality QA/RC data is generally scarce (see §5.1).

Documented data. Much work has been invested in investigation of biases in the current resources (see §7.1), and the conclusion is clear: if the data has statistically conspicuous "shortcuts", we have no reason to expect neural nets to not pick up on them [154]. Much work discussed in this survey proposed various improvements to the data collection methodology (which deserves a separate survey), but it is hard to guarantee absence of spurious patterns in naturally-occurring data – and it gets harder as dataset size grows [87]. The field of AI ethics is currently working on documenting the speaker demographics and possible social biases [21, 88], with the idea that this would then be useful for model certification [178]. Given that neither social nor linguistic spurious patterns in the data are innocuous [219], we need a similar practice for documenting spurious patterns in the data we use. New datasets will be a lot more useful if they come with documented limitations, rather than with the impression that there are none.

8 CONCLUSION

The number of QA/RC datasets produced by the NLP community is large and growing rapidly. We have presented the most extensive survey of the field to date, identifying the key dimensions along which the current datasets vary. These dimensions provide a conceptual framework for evaluating current and future resources in terms of their format, domain, and target reasoning skills. We have categorized over two hundred datasets while highlighting the gaps in the current literature, and we hope that this survey would be useful both for the NLP practitioners looking for data, and for those seeking to push the boundaries of QA/RC research.

REFERENCES

[1] Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard. 2019. X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. ACL, Hong Kong, China, 265–274. https://doi.org/10.18653/v1/D19-6130

[2] Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. ComQA: A Community-Sourced Dataset for Complex Factoid Question Answering with Paraphrase Clusters. In *NAACL-HLT*. 307–317. https://aclweb.org/anthology/papers/N/N19/N19-1027/

[3] Manoj Acharya, Karan Jariwala, and Christopher Kanan. 2019. VQD: Visual Query Detection In Natural Scenes. In *NAACL-HLT*. ACL, Minneapolis, Minnesota, 1955–1961. https://doi.org/10.18653/v1/N19-1194

[4] Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering Complex Counting Questions. In *AAAI*. arXiv:1810.12440 http://arxiv.org/abs/1810.12440

[5] Douglas Adams. 2009. *The Hitchhiker's Guide to the Galaxy* (del rey trade pbk. ed ed.). Ballantine Books, New York.

[6] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. *Dialogue Acts in Verbmobil 2.* Technical Report. Verbmobil.

[7] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *NAACL*. ACL, Online, 520–534. https://aclanthology.org/2021.naacl-main.44

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, 2425–2433. http://ieeexplore.ieee.org/document/7410636/

[9] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-Lingual Transferability of Monolingual Representations. In *ACL*. ACL, 4623–4637. https://doi.org/10.18653/v1/2020.acl-main.421 arXiv:1910.11856

[10] Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual Extractive Reading Comprehension by Runtime Machine Translation. *arXiv:1809.03275 [cs]* (nov 2018). arXiv:cs/1809.03275 http://arxiv.org/abs/1809.03275

[11] Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. XOR QA: Cross-Lingual Open-Retrieval Question Answering. *arXiv:2010.11856 [cs]* (oct 2020). arXiv:cs/2010.11856 http://arxiv.org/abs/2010.11856

[12] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. *arXiv:1704.00057 [cs]* (apr 2017). arXiv:cs/1704.00057 http://arxiv.org/abs/1704.00057

[13] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[14] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4685–4697. https://doi.org/10.18653/v1/D19-1475

[15] Kathleen M. Bailey. 2018. Multiple-Choice Item Format. *The TESOL Encyclopedia of English Language Teaching* (jan 2018), 1–8. https://doi.org/10.1002/9781118784235.eelt0369

[16] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268 [cs]* (nov 2016). arXiv:cs/1611.09268 http://arxiv.org/abs/1611.09268

[17] Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2017. Embracing Data Abundance: BookTest Dataset for Reading Comprehension. In *ICLR*. arXiv:1610.00956 https://openreview.net/pdf?id=H1U4mhVFe

[18] Somnath Banerjee, Sudip Kumar Naskar, and Paolo Rosso. [n.d.]. The First Cross-Script Code-Mixed Question Answering Corpus. ([n. d.]), 10. http://ceur-ws.org/Vol-1589/MultiLingMine6.pdf

[19] Paul Bartha. 2019. Analogy and Analogical Reasoning. In *The Stanford Encyclopedia of Philosophy* (spring 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/reasoning-analogy/

[20] Emily M. Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

[21] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *TACL* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[22] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, Seattle, Washington, USA, 1533–1544. https://www.aclweb.org/anthology/D13-1160

[23] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling Biological Processes for Reading Comprehension. In *EMNLP (EMNLP)*. 1499–1510.

[24] Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured Annotations for Reading Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Online, 5726–5735. arXiv:2004.14797 https://www.aclweb.org/anthology/2020.acl-main.507

[25] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Byg1v1HKDB

[26] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *EMNLP (EMNLP)*. ACL, Online, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

[27] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*. arXiv:1911.11641 http://arxiv.org/abs/1911.11641

[28] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *EMNLP (EMNLP)*. ACL, Online, 5480–5494. https://doi.org/10.18653/v1/2020.emnlp-main.442

[29] Simon Blackburn. 2008. Inference. https://www.oxfordreference.com/view/10.1093/acref/9780199541430.001.0001/acref-9780199541430

[30] Simon Blackburn. 2008. Reasoning. https://www.oxfordreference.com/view/10.1093/acref/9780199541430.001.0001/acref-9780199541430

[31] Elisa Bone and Mike Prosser. 2020. Multiple Choice Questions: An Introductory Guide. (2020). https://melbourne-cshe.unimelb.edu.au/__data/assets/pdf_file/0010/3430648/multiple-choice-questions_final.pdf

[32] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-Scale Simple Question Answering with Memory Networks. *arXiv:1506.02075 [cs]* (jun 2015). arXiv:cs/1506.02075 http://arxiv.org/abs/1506.02075

[33] Jordan Boyd-Graber. 2019. What Question Answering Can Learn from Trivia Nerds. *arXiv:1910.14464 [cs]* (oct 2019). arXiv:cs/1910.14464 http://arxiv.org/abs/1910.14464

[34] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. *arXiv:2005.14165 [cs]* (jun 2020). arXiv:cs/2005.14165 http://arxiv.org/abs/2005.14165

[35] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 5016–5026. https://doi.org/10.18653/v1/D18-1547

[36] Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA-Accessing Domain-Specific FAQs via Conversational QA. In *ACL*. ACL, Online, 7302–7314.

[37] Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering. *arXiv:1912.05200 [cs]* (dec 2019). arXiv:cs/1912.05200 http://arxiv.org/abs/1912.05200

[38] Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Michael McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avi Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang. 2020. The TechQA Dataset. In *ACL*. ACL, Online, 1269–1278. https://www.aclweb.org/anthology/2020.acl-main.117

[39] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of Text Generation: A Survey. *arXiv:2006.14799 [cs]* (jun 2020). arXiv:cs/2006.14799 http://arxiv.org/abs/2006.14799

[40] Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W. Black. 2018. Code-Mixed Question Answering Challenge: Crowd-Sourcing Data and Techniques. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. ACL, Melbourne, Australia, 29–38. https://doi.org/10.18653/v1/W18-3204

[41] Graham Chapman, John Cleese, Terry Gilliam, Eric Idle, Terry Jones, Michael Palin, John Goldstone, Spike Milligan, Monty Python (Comedy troupe), Handmade Films, and Criterion Collection (Firm). 1999. Life of Brian.

[42] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. 2017. Counting Everyday Objects in Everyday Scenes. In *CVPR*. 1135–1144. https://openaccess.thecvf.com/content_cvpr_2017/html/Chattopadhyay_Counting_Everyday_Objects_CVPR_2017_paper.html

[43] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating Question Answering Evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. ACL, Hong Kong, China, 119–124. https://doi.org/10.18653/v1/D19-5817

[44] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A Dataset for Training and Evaluating Generative Reading Comprehension Metrics. In *EMNLP (EMNLP)*. ACL, Online, 6521–6532. https://www.aclweb.org/anthology/2020.emnlp-main.528

[45] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *ACL*. ACL, Vancouver, Canada, 1870–1879. https://doi.org/10.18653/v1/P17-1171

[46] Danqi Chen and Wen-tau Yih. 2020. Open-Domain Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. ACL, Online, 34–37. https://doi.org/10.18653/v1/2020.acl-tutorials.8

[47] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. ACL, Online, 1026–1036. https://doi.org/10.18653/v1/2020.findings-emnlp.91

[48] Carlos Iván Chesnevar, Ana Gabriela Maguitman, and Ronald Prescott Loui. 2000. Logical Models of Argument. *ACM Computing Surveys (CSUR)* 32, 4 (2000), 337–383. https://dl.acm.org/doi/pdf/10.1145/371578.371581

[49] Minseok Cho, Reinald Kim Amplayo, Seung-won Hwang, and Jonghyuck Park. 2018. Adversarial TableQA: Attention Supervision for Question Answering on Tables. In *Proceedings of Machine Learning Research*. 391–406. http://proceedings.mlr.press/v95/cho18a/cho18a.pdf

[50] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *EMNLP*. ACL, Brussels, Belgium, 2174–2184. http://aclweb.org/anthology/D18-1241

[51] Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *TACL* 9 (apr 2021), 447–461. https://doi.org/10.1162/tacl_a_00377

[52] Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *ACL*. ACL, Melbourne, Australia, 845–855. https://doi.org/10.18653/v1/P18-1078

[53] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL-HLT*. 2924–2936. https://aclweb.org/anthology/papers/N/N19/N19-1300/

[54] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *TACL* 8 (jul 2020), 454–470. https://doi.org/10.1162/tacl_a_00317

[55] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457 [cs]* (mar 2018). arXiv:cs/1803.05457 http://arxiv.org/abs/1803.05457

[56] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang, and Doo Soon Kim. 2020. TutorialVQA: Question Answering Dataset for Tutorial Videos. In *LREC*. ELRA, Marseille, France, 5450–5455. https://www.aclweb.org/anthology/2020.lrec-1.670

[57] Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. *arXiv:2004.11861 [cs]* (apr 2020). arXiv:cs/2004.11861 http://arxiv.org/abs/2004.11861

[58] Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2019. Enabling Deep Learning for Large Scale Question Answering in Italian. *Intelligenza Artificiale* 13, 1 (jan 2019), 49–61. https://doi.org/10.3233/IA-190018

[59] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 5883–5889. https://doi.org/10.18653/v1/D19-1600

[60] Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the First Evaluation on Chinese Machine Reading Comprehension. In *LREC*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1431

[61] Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus Attention-Based Neural Networks for Chinese Reading Comprehension. In *COLING*. The COLING 2016 Organizing Committee, Osaka, Japan, 1777–1786. https://www.aclweb.org/anthology/C16-1167

[62] Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. 2020. A Sentence Cloze Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*. ICLR, Barcelona, Spain (Online), 6717–6723. https://doi.org/10.18653/v1/2020.coling-main.589

[63] Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 5924–5931. https://doi.org/10.18653/v1/D19-1606

[64] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *NAACL-HLT*. ACL, Online, 4599–4610. https://doi.org/10.18653/v1/2021.naacl-main.365

[65] Lorenz Demey, Barteld Kooi, and Joshua Sack. 2019. Logic and Probability. In *The Stanford Encyclopedia of Philosophy* (summer 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/logic-probability/

[66] Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. FQuAD: French Question Answering Dataset. *arXiv:2002.06071 [cs]* (feb 2020). arXiv:cs/2002.06071 http://arxiv.org/abs/2002.06071

[67] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. *arXiv:1811.01241 [cs]* (nov 2018). arXiv:cs/1811.01241 http://arxiv.org/abs/1811.01241

[68] Cícero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning Hybrid Representations to Retrieve Semantically Equivalent Questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, 694–699. https://doi.org/10.3115/v1/P15-2114

[69] Igor Douven. 2017. Abduction. In *The Stanford Encyclopedia of Philosophy* (summer 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/abduction/

[70] Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Matt Gardner, and Sameer Singh. 2019. ORB: An Open Reading Benchmark for Comprehensive Evaluation of Machine Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. ACL, Hong Kong, China, 147–153. https://doi.org/10.18653/v1/D19-5820

[71] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL-HLT*. 2368–2378. https://aclweb.org/anthology/papers/N/N19/N19-1246/

[72] Nan Duan and Duyu Tang. 2018. Overview of the NLPCC 2017 Shared Task: Open Domain Chinese Question Answering. In *Natural Language Processing and Chinese Computing*, Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong (Eds.). Springer, Cham, 954–961. https://doi.org/10.1007/978-3-319-73618-1_86

[73] Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To Test Machine Comprehension, Start by Defining Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Online, 7839–7859. https://www.aclweb.org/anthology/2020.acl-main.701

[74] Daria Dzendzik, Carl Vogel, and Jennifer Foster. 2021. English Machine Reading Comprehension Datasets: A Survey. *arXiv:2101.10421 [cs]* (jan 2021). arXiv:cs/2101.10421 http://arxiv.org/abs/2101.10421

[75] Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. *arXiv:1912.09723 [cs]* (may 2020). https://doi.org/10.1007/978-3-030-58219-7_1 arXiv:cs/1912.09723

[76] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 5918–5924. https://doi.org/10.18653/v1/D19-1605

[77] Mihail Eric and Christopher D. Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. *arXiv:1705.05414 [cs]* (jul 2017). arXiv:cs/1705.05414 http://arxiv.org/abs/1705.05414

[78] Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *TACL* 8 (2020), 34–48. https://doi.org/10.1162/tacl_a_00298 arXiv:1907.13528

[79] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *ACL*. ACL, Florence, Italy, 3558–3567. https://doi.org/10.18653/v1/P19-1346

[80] H. M. Fayek and J. Johnson. 2020. Temporal Reasoning via Audio Question Answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2283–2294. https://doi.org/10.1109/TASLP.2020.3010650

[81] Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. 2020. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*. ICLR, Barcelona, Spain (Online), 6481–6497. https://doi.org/10.18653/v1/2020.coling-main.570

[82] James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. 2020. IIRC: A Dataset of Incomplete Information Reading Comprehension Questions. In *EMNLP*. ACL, Online, 1137–1147. https://doi.org/10.18653/v1/2020.emnlp-main.86

[83] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. ACL, Hong Kong, China, 1–13. https://doi.org/10.18653/v1/D19-5801

[84] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A Dataset and Baselines for Visual Question Answering on Art. *arXiv:2008.12520 [cs]* (aug 2020). arXiv:cs/2008.12520 http://arxiv.org/abs/2008.12520

[85] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. ACL, Online, 1307–1323. https://doi.org/10.18653/v1/2020.findings-emnlp.117

[86] Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question Answering Is a Format; When Is It Useful? *arXiv:1909.11291 [cs]* (sep 2019). arXiv:cs/1909.11291 http://arxiv.org/abs/1909.11291

[87] Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency Problems: On Finding and Removing Artifacts in Language Data. *arXiv:2104.08646 [cs]* (April 2021). arXiv:cs/2104.08646 http://arxiv.org/abs/2104.08646

[88] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2020). arXiv:cs/1803.09010 http://arxiv.org/abs/1803.09010

[89] Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing Fair Generalization Tasks for Natural Language Inference. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 4475–4485. https://doi.org/10.18653/v1/D19-1456

[90] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *EMNLP-IJCNLP*. ACL, Hong Kong, China, 1161–1166. https://doi.org/10.18653/v1/D19-1107

[91] Taisia Glushkova, Alexey Machnev, Alena Fenogenova, Tatiana Shavrina, Ekaterina Artemova, and Dmitry I. Ignatov. 2020. DaNetQA: A Yes/No Question Answering Dataset for the Russian Language. *arXiv:2010.02605 [cs]* (oct 2020). arXiv:cs/2010.02605 http://arxiv.org/abs/2010.02605

[92] Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv:1901.05287 [cs]* (jan 2019). arXiv:cs/1901.05287 http://arxiv.org/abs/1901.05287

[93] Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the Interaction of Belief Bias and Explanations. (Aug. 2021), 2930–2942. https://aclanthology.org/2021.findings-acl.259

[94] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. 394–398. https://aclweb.org/anthology/papers/S/S12/S12-1052/

[95] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual Question Answering in Interactive Environments. In *CVPR*. 4089–4098.

[96] Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. *arXiv:2011.07743 [cs]* (feb 2021). https://doi.org/10.1145/3442381.3449992 arXiv:cs/2011.07743

[97] Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2020. MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models. *arXiv:2005.02507 [cs]* (may 2020). arXiv:cs/2005.02507 http://arxiv.org/abs/2005.02507

[98] Shangmin Guo, Kang Liu, Shizhu He, Cao Liu, Jun Zhao, and Zhuoyu Wei. 2017. IJCNLP-2017 Task 5: Multi-Choice Question Answering in Examinations. In *Proceedings of the IJCNLP 2017, Shared Tasks*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 34–40. https://www.aclweb.org/anthology/I17-4005

[99] Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering. In *Findings of ACL*. https://arxiv.org/abs/2106.04016

[100] Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A Multi-Domain Multi-Lingual Question-Answering Framework for English and Hindi. In *LREC*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1440

[101] Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. AmazonQA: A Review-Based Question Answering Task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macao, China, 4996–5002. https://doi.org/10.24963/ijcai.2019/694

[102] Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava. 2018. Transliteration Better than Translation? Answering Code-Mixed Questions over a Knowledge Base. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. ACL, Melbourne, Australia, 39–50. https://doi.org/10.18653/v1/W18-3205

[103] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL-HLT*. ACL, New Orleans, Louisiana, 107–112. https://doi.org/10.18653/v1/N18-2017

[104] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2019. ANTIQUE: A Non-Factoid Question Answering Benchmark. *arXiv:1905.08957 [cs]* (Aug. 2019). arXiv:cs/1905.08957 http://arxiv.org/abs/1905.08957

[105] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster.. In *KDD*. ACM, 1803–1812. http://dblp.uni-trier.de/db/conf/kdd/kdd2017.html#HassanALT17

[106] James Hawthorne. 2021. Inductive Logic. In *The Stanford Encyclopedia of Philosophy* (spring 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/logic-inductive/

[107] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: A Chinese Machine Reading Comprehension Dataset from Real-World Applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*. ACL, Melbourne, Australia, 37–46. https://doi.org/10.18653/v1/W18-2605

[108] Nancy Hedberg, Juan M Sosa, and Lorna Fadden. 2004. Meanings and configurations of questions in English. In *Speech Prosody 2004, International Conference*.

[109] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*. https://www.aclweb.org/anthology/H90-1021

[110] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 1693–1701. http://dl.acm.org/citation.cfm?id=2969239.2969428

[111] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A Novel Large-Scale Language Understanding Task over Wikipedia. In *ACL*. ACL, Berlin, Germany, 1535–1545. https://doi.org/10.18653/v1/P16-1145

[112] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]* (nov 2015). arXiv:cs/1511.02301 http://arxiv.org/abs/1511.02301

[113] Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. Linguistic Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions. In *LREC*. ELRA, Marseille, France, 1753–1762. https://www.aclweb.org/anthology/2020.lrec-1.217

[114] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 2391–2401. https://doi.org/10.18653/v1/D19-1243

[115] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*. 6700–6709. https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html

[116] SRI International. 2011. SRI's Amex Travel Agent Data. http://www.ai.sri.com/~communic/amex/amex.html

[117] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*. 2758–2766. https://openaccess.thecvf.com/content_cvpr_2017/html/Jang_TGIF-QA_Toward_Spatio-Temporal_CVPR_2017_paper.html

[118] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *EMNLP*. ACL, 2021–2031. https://doi.org/10.18653/v1/D17-1215

[119] Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. TempQuestions: A Benchmark for Temporal Question Answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. ACM Press, Lyon, France, 1057–1062. https://doi.org/10.1145/3184558.3191536

[120] Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, Torino, Italy, 1807–1810. https://doi.org/10.1145/3269206.3269247

[121] Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase. In *NAACL-HLT*. 318–323. https://aclweb.org/anthology/papers/N/N19/N19-1028/

[122] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 2567–2577. https://doi.org/10.18653/v1/D19-1259

[123] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2901–2910.

[124] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*. ACL, 1601–1611. https://doi.org/10.18653/v1/P17-1147

[125] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*. arXiv:1909.12434 https://openreview.net/forum?id=Sklgs0NFvr

[126] Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a Native French Question-Answering Dataset. In *LREC*. ELRA, Marseille, France, 5481–5490. https://www.aclweb.org/anthology/2020.lrec-1.673

[127] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *NAACL-HLT*. 252–262. https://doi.org/10.18653/v1/N18-1023

[128] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System. *arXiv:2005.00700 [cs]* (2020). arXiv:cs/2005.00700 https://arxiv.org/abs/2005.00700

[129] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. DeepStory: Video Story QA by Deep Embedded Memory Networks. In *IJCAI*. https://openreview.net/forum?id=ryZczSz_bS

[130] Seokhwan Kim, Luis Ferdinando D'Haro, Rafael E. Banchs, Matthew Henderson, Jason Willisams, and Koichiro Yoshino. 2016. Dialog State Tracking Challenge 5 Handbook v.3.1. http://workshop.colips.org/dstc5/

[131] Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *TACL* 6 (2018), 317–328. http://aclweb.org/anthology/Q18-1023

[132] Xiang Kong, Varun Gangal, and Eduard Hovy. 2020. SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations. In *ACL*. ACL, Online, 5668–5683. https://doi.org/10.18653/v1/2020.acl-main.502

[133] Robert Koons. 2017. Defeasible Reasoning. In *The Stanford Encyclopedia of Philosophy* (winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/reasoning-defeasible/

[134] Vladislav Korablinov and Pavel Braslavski. 2020. RuBQ: A Russian Dataset for Question Answering over Wikidata. *arXiv:2005.10659 [cs]* (may 2020). arXiv:cs/2005.10659 http://arxiv.org/abs/2005.10659

[135] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-Form Question Answering. In *NAACL-HLT*. ACL, Online, 4940–4957. https://doi.org/10.18653/v1/2021.naacl-main.393

[136] Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to Automatically Solve Algebra Word Problems. In *ACL*. ACL, Baltimore, Maryland, 271–281. https://doi.org/10.3115/v1/P14-1026

[137] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics* (2019). https://ai.google/research/pubs/pub47761

[138] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-Scale ReAding Comprehension Dataset From Examinations. In *EMNLP*. ACL, 785–794. https://doi.org/10.18653/v1/D17-1082

[139] C. Lee, S. Wang, H. Chang, and H. Lee. 2018. ODSQA: Open-Domain Spoken Question Answering Dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. 949–956. https://doi.org/10.1109/SLT.2018.8639505

[140] Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-Supervised Training Data Generation for Multilingual Question Answering. In *LREC*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1437

[141] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*. 1369–1379. https://www.aclweb.org/anthology/papers/D/D18/D18-1167/

[142] Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. 552–561.

[143] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. ACL, Vancouver, Canada, 333–342. https://doi.org/10.18653/v1/K17-1034

[144] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-Lingual Extractive Question Answering. In *ACL*. ACL, Online, 7315–7330. https://www.aclweb.org/anthology/2020.acl-main.653/

[145] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. *arXiv:1804.00320 [cs]* (apr 2018). arXiv:cs/1804.00320 http://arxiv.org/abs/1804.00320

[146] Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A Challenge Multiparty Dialogues-Based Machine Reading Comprehension Dataset with Discourse Structure. *arXiv:2004.05080 [cs]* (apr 2020). arXiv:cs/2004.05080 http://arxiv.org/abs/2004.05080

[147] Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. *arXiv:1607.06275 [cs]* (sep 2016). arXiv:cs/1607.06275 http://arxiv.org/abs/1607.06275

[148] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. In *ACL*. ACL, Florence, Italy, 1340–1350. https://doi.org/10.18653/v1/P19-1129

[149] Yichan Liang, Jianheng Li, and Jian Yin. 2019. A New Multi-Choice Reading Comprehension Dataset for Curriculum Learning. In *Proceedings of the Eleventh Asian Conference on Machine Learning (Proceedings of Machine Learning Research)*, Wee Sun Lee and Taiji Suzuki (Eds.), Vol. 101. PMLR, Nagoya, Japan, 742–757. http://proceedings.mlr.press/v101/liang19a.html

[150] Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension. *arXiv:1909.07005 [cs]* (sep 2019). arXiv:cs/1909.07005 http://arxiv.org/abs/1909.07005

[151] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *EMNLP (EMNLP)*. ACL, Online, 6862–6868. https://doi.org/10.18653/v1/2020.emnlp-main.557

[152] Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning Over Paragraph Effects in Situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. ACL, Hong Kong, China, 58–62. https://doi.org/10.18653/v1/D19-5808

[153] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[154] Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *arXiv:2005.00955 [cs]* (may 2020). arXiv:cs/2005.00955 https://arxiv.org/pdf/2005.00955.pdf

[155] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning.. In *IJCAI*, Christian Bessiere (Ed.). ijcai.org, 3622–3628. http://dblp.uni-trier.de/db/conf/ijcai/ijcai2020.html#LiuCLHWZ20 Scheduled for July 2020, Yokohama, Japan, postponed due to the Corona pandemic.

[156] Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A Cross-Lingual Open-Domain Question Answering Dataset. In *ACL*. ACL, Florence, Italy, 2358–2368. https://doi.org/10.18653/v1/P19-1227

[157] Pengyuan Liu, Yuning Deng, Chenghao Zhu, and Han Hu. 2019. XCMRC: Evaluating Cross-Lingual Machine Reading Comprehension. In *Natural Language Processing and Chinese Computing*, Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan (Eds.). Springer, Cham, 552–564.

https://doi.org/10.1007/978-3-030-32233-5_43

[158] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]* (July 2019). arXiv:cs/1907.11692 http://arxiv.org/abs/1907.11692

[159] Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World Knowledge for Reading Comprehension: Rare Entity Prediction with Hierarchical LSTMs Using External Descriptions. In *EMNLP*. ACL, 825–834. https://doi.org/10.18653/v1/D17-1086

[160] Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *arXiv:2007.15207 [cs]* (jul 2020). arXiv:cs/2007.15207 http://arxiv.org/abs/2007.15207

[161] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *arXiv:1506.08909 [cs]* (jun 2015). arXiv:cs/1506.08909 http://arxiv.org/abs/1506.08909

[162] Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging Reading Comprehension on Daily Conversation: Passage Completion on Multiparty Dialog. In *NAACL*. ACL, New Orleans, Louisiana, 2039–2048. https://doi.org/10.18653/v1/N18-1185

[163] Leigh-Ann MacFarlane and Genvième Boulet. 2017. Multiple-Choice Tests Can Support Deep Learning! *Proceedings of the Atlantic Universities' Teaching Showcase* 21, 0 (2017), 61–66. https://ojs.library.dal.ca/auts/article/view/8430

[164] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A Dataset and Exploration of Models for Understanding Video Data through Fill-in-the-Blank Question-Answering. *arXiv:1611.07810 [cs]* (feb 2017). arXiv:cs/1611.07810 http://arxiv.org/abs/1611.07810

[165] Cheryl L. Marcham, Treasa M. Turnbeaugh, Susan Gould, and Joel T. Nadler. 2018. Developing Certification Exam Questions: More Deliberate Than You May Think. *Professional Safety* 63, 05 (may 2018), 44–49. https://onepetro.org/PS/article/63/05/44/33528/Developing-Certification-Exam-Questions-More

[166] Michał Marcinczuk, Marcin Ptak, Adam Radziszewski, and Maciej Piasecki. 2013. Open Dataset for Development of Polish Question Answering Systems. In *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Wydawnictwo Poznanskie, Fundacja Uniwersytetu im. Adama Mickiewicza. https://www.researchgate.net/profile/Maciej-Piasecki/publication/272685856_Open_dataset_for_development_of_Polish_Question_Answering_systems/links/560f8ff708aec422d1133caa/Open-dataset-for-development-of-Polish-Question-Answering-systems.pdf

[167] Julian McAuley and Alex Yang. 2016. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. In *WWW (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 625–635. https://doi.org/10.1145/2872427.2883044

[168] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv:1806.08730 [cs, stat]* (jun 2018). arXiv:cs, stat/1806.08730 http://arxiv.org/abs/1806.08730

[169] John McCarthy and Patrick Hayes. 1969. Some Philosophical Problems From the Standpoint of Artificial Intelligence. In *Machine Intelligence 4*, B. Meltzer and Donald Michie (Eds.). Edinburgh University Press, 463–502.

[170] R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. BERTs of a Feather Do Not Generalize Together: Large Variability in Generalization across Models with Similar Test Set Performance. *arXiv:1911.02969 [cs]* (nov 2019). arXiv:cs/1911.02969 http://arxiv.org/abs/1911.02969

[171] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, Florence, Italy, 3428–3448. https://doi.org/10.18653/v1/P19-1334

[172] Danielle S. McNamara and Joe Magliano. 2009. Chapter 9 Toward a Comprehensive Model of Comprehension. In *Psychology of Learning and Motivation*. The Psychology of Learning and Motivation, Vol. 51. Academic Press, 297–384. https://doi.org/10.1016/S0079-7421(09)51009-2

[173] Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 975–984. https://www.aclweb.org/anthology/2020.acl-main.92

[174] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*. ACL, Brussels, Belgium, 2381–2391. http://aclweb.org/anthology/D18-1260

[175] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-Domain Questions. *arXiv:2004.10645 [cs]* (apr 2020). arXiv:cs/2004.10645 http://arxiv.org/abs/2004.10645

[176] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *NAACL*. ACL, Online, 4582–4598. https://doi.org/10.18653/v1/2021.naacl-main.364

[177] Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, and Chitta Baral. 2020. Towards Question Format Independent Numerical Reasoning: A Set of Prerequisite Tasks. *arXiv preprint arXiv:2005.08516* (2020). arXiv:2005.08516 https://arxiv.org/abs/2005.08516

[178] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *FCccT (FAT* '19)*. ACM, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[179] Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 3485–3493. https://www.aclweb.org/anthology/L16-1555

[180] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A Question Answering Dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. ACL, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.18

[181] Nasrin Mostafazadeh, Michael Roth, Nathanael Chambers, and Annie Louis. 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*. ACL, 46–51. http://www.aclweb.org/anthology/W17-0900

[182] Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic Question Answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. ACL, Florence, Italy, 108–118. https://doi.org/10.18653/v1/W19-4612

[183] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. MarioQA: Answering Questions by Watching Gameplay Videos. In *ICCV*. arXiv:1612.01669 http://arxiv.org/abs/1612.01669

[184] Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*. Vancouver, Canada, August 3 - 4, 2017, 27–48. http://www.aclweb.org/anthology/S17-2003

[185] Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. SemEval-2015 Task 3: Answer Selection in Community Question Answering. 269–281.

[186] Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community Question Answering. 525–545.

[187] Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese Dataset for Evaluating Machine Reading Comprehension. In *ICLR*. ICLR, Barcelona, Spain (Online), 2595–2605. https://doi.org/10.18653/v1/2020.coling-main.233

[188] Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. *arXiv:2005.00242 [cs]* (may 2020). arXiv:cs/2005.00242 http://arxiv.org/abs/2005.00242

[189] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. 2020. A Method for Building a Commonsense Inference Dataset Based on Basic Events. In *EMNLP (EMNLP)*. ACL, Online, 2450–2460. https://www.aclweb.org/anthology/2020.emnlp-main.192

[190] Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who Did What: A Large-Scale Person-Centered Cloze Dataset. In *EMNLP*. ACL, Austin, Texas, 2230–2235. https://doi.org/10.18653/v1/D16-1241

[191] Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 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)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1564

[192] Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. ACL, New Orleans, Louisiana, 747–757. https://doi.org/10.18653/v1/S18-1119

[193] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *EMNLP*. Brussels, Belgium, 2357–2368. http://aclweb.org/anthology/D18-1258

[194] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA Dataset: Word Prediction Requiring a Broad Discourse Context. *arXiv:1606.06031 [cs]* (2016). arXiv:cs/1606.06031 http://arxiv.org/abs/1606.06031

[195] Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *ACL-IJCNLP*. ACL, Beijing, China, 1470–1480. https://doi.org/10.3115/v1/P15-1142

[196] Alkesh Patel, Akanksha Bindal, Hadas Kotek, Christopher Klein, and Jason Williams. 2020. Generating Natural Questions from Images for Multimodal Assistants. *arXiv:2012.03678 [cs]* (nov 2020). arXiv:cs/2012.03678 http://arxiv.org/abs/2012.03678

[197] Anselmo Peñas, Christina Unger, and Axel-Cyrille Ngonga Ngomo. 2014. Overview of CLEF Question Answering Track 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. Springer, Cham, 300–306. https://doi.org/10.1007/978-3-319-11382-1_23

[198] Anselmo Peñas, Christina Unger, Georgios Paliouras, and Ioannis Kakadiaris. 2015. Overview of the CLEF Question Answering Track 2015. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (Lecture Notes in Computer Science)*. 539–544.

[199] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: A Novel Multi-Domain Information Seeking Dialogues Dataset. *arXiv:1912.04639 [cs]* (dec 2019). arXiv:cs/1912.04639 http://arxiv.org/abs/1912.04639

[200] Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies toward Curating and Annotating Large Scale Dialogue Data. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 4526–4536. https://doi.org/10.18653/v1/D19-1460

[201] Eric Price. 2014. The NIPS Experiment. http://blog.mrtz.org/2014/12/15/the-nips-experiment.html

[202] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2019. Learning to Deceive with Attention-Based Explanations. *arXiv:1909.07913 [cs]* (sep 2019). arXiv:cs/1909.07913 http://arxiv.org/abs/1909.07913

[203] Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. arXiv:cs.CL/2106.04571

[204] Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A Survey on Neural Machine Reading Comprehension. *arXiv:1906.03824 [cs]* (June 2019). arXiv:cs/1906.03824 http://arxiv.org/abs/1906.03824

[205] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-Seeking Conversations. In *SIGIR (SIGIR '18)*. ACM, New York, NY, USA, 989–992. https://doi.org/10.1145/3209978.3210124

[206] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. (2019). https://research.google/pubs/pub48414/

[207] Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. "Answer Ka Type Kya He?": Learning to Classify Questions in Code-Mixed Language. In *WWW (WWW '15 Companion)*. ACM, New York, NY, USA, 853–858. https://doi.org/10.1145/2740908.2743006

[208] Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. 2020. Explaining and Improving Model Behavior with k Nearest Neighbor Representations. *arXiv:2010.09030 [cs]* (oct 2020). arXiv:cs/2010.09030 http://arxiv.org/abs/2010.09030

[209] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *ACL*. ACL, Melbourne, Australia, 784–789. http://aclweb.org/anthology/P18-2124

[210] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*. 2383–2392.

[211] Alan Ramponi and Barbara Plank. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. ICLR, Barcelona, Spain (Online), 6838–6855. https://doi.org/10.18653/v1/2020.coling-main.603

[212] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *ACL*. ACL, Melbourne, Australia, 463–473. https://doi.org/10.18653/v1/P18-1043

[213] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *TACL* 7 (mar 2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[214] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, San Francisco, California, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[215] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Online, 4902–4912. https://www.aclweb.org/anthology/2020.acl-main.442

[216] Matthew Richardson, Christopher J C Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP*. Seattle, Washington, USA, 18-21 October 2013, 193–203.

[217] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium: Logical For- Malizations of Commonsense Reasoning*. 6.

[218] Anna Rogers. 2019. How the Transformers Broke NLP Leaderboards. https://hackingsemantics.xyz/2019/leaderboards/

[219] Anna Rogers. 2021. Changing the World by Changing the Data. In *ACL*. ACL, Online, 2182–2194. https://aclanthology.org/2021.acl-long.170

[220] Anna Rogers and Isabelle Augenstein. 2020. What Can We Do to Improve Peer Review in NLP?. In *Findings of EMNLP*. ACL, Online, 1256–1262. https://www.aclweb.org/anthology/2020.findings-emnlp.112/

[221] Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *AAAI*. 8722–8731. https://aaai.org/ojs/index.php/AAAI/article/view/6398

[222] Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool. *arXiv:2004.05484 [cs]* (apr 2020). arXiv:cs/2004.05484 http://arxiv.org/abs/2004.05484

[223] Sebastian Ruder and Si Avirup. 2021. Multi-Domain Multilingual Question Answering. In *EMNLP*.

[224] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking Like a Skeptic: Defeasible Inference in Natural Language. In *Findings of EMNLP 2020*. ACL, Online, 4661–4675. https://doi.org/10.18653/v1/2020.findings-emnlp.418

[225] Barbara Rychalska, Dominika Basaj, Anna Wróblewska, and Przemyslaw Biecek. 2018. Does It Care What You Asked? Understanding Importance of Verbs in Deep Learning QA System. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*. ACL, 322–324. http://aclweb.org/anthology/W18-5436

[226] Mrinmaya Sachan, Kumar Dubey, Eric Xing, and Matthew Richardson. 2015. Learning Answer-Entailing Structures for Machine Comprehension. In *ACL-IJCNLP*. ACL, Beijing, China, 239–249. https://doi.org/10.3115/v1/P15-1024

[227] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *EMNLP*. ACL, Brussels, Belgium, 2087–2097. https://doi.org/10.18653/v1/D18-1233

[228] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. *arXiv:1907.10641 [cs]* (jul 2019). arXiv:cs/1907.10641 http://arxiv.org/abs/1907.10641

[229] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 4453–4463. https://doi.org/10.18653/v1/D19-1454

[230] Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2020. Beyond Leaderboards: A Survey of Methods for Revealing Weaknesses in Natural Language Inference Data and Models. *arXiv:2005.14709 [cs]* (May 2020). arXiv:cs/2005.14709 http://arxiv.org/abs/2005.14709

[231] Viktor Schlegel, Marco Valentino, André Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A Framework for Evaluation of Machine Reading Comprehension Gold Standards. In *Language Resources and Evaluation Conference*. arXiv:2003.04642 http://arxiv.org/abs/2003.04642

[232] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *arXiv:1512.05742 [cs, stat]* (dec 2015). arXiv:cs, stat/1512.05742 http://arxiv.org/abs/1512.05742

[233] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. DRCD: A Chinese Machine Reading Comprehension Dataset. *arXiv:1806.00920 [cs]* (may 2019). arXiv:cs/1806.00920 http://arxiv.org/abs/1806.00920

[234] Shuming Shi, Yuehui Wang, Chin-Yew Lin, Xiaojiang Liu, and Yong Rui. 2015. Automatically Solving Number Word Problems by Semantic Parsing and Reasoning. In *EMNLP*. ACL, Lisbon, Portugal, 1132–1142. https://doi.org/10.18653/v1/D15-1135

[235] Hideyuki Shibuki, Kotaro Sakamoto, Yoshionobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab Task. *Proceedings of the 11th NTCIR Conference* (2014), 518–529. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-QALAB-ShibukiH.pdf

[236] Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In *EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 4496–4505. https://doi.org/10.18653/v1/D19-1458

[237] Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. 2020. A Survey of Code-Switched Speech and Language Processing. *arXiv:1904.00784 [cs, stat]* (jul 2020). arXiv:cs, stat/1904.00784 http://arxiv.org/abs/1904.00784

[238] Amir Soleimani, Christof Monz, and Marcel Worring. 2021. NLQuAD: A Non-Factoid Long Question Answering Data Set. In *EACL*. ACL, Online, 1245–1255. https://aclanthology.org/2021.eacl-main.106

[239] Saku Sugawara and Akiko Aizawa. 2016. An Analysis of Prerequisite Skills for Reading Comprehension. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*. ACL, Austin, TX, 1–5. https://doi.org/10.18653/v1/W16-6001

[240] Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability. In *ACL*. ACL, Vancouver, Canada, 806–817. https://doi.org/10.18653/v1/P17-1075

[241] Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In *AAAI*. arXiv:1911.09241 http://arxiv.org/abs/1911.09241

[242] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A Corpus of Natural Language for Visual Reasoning. In *ACL*. 217–223. https://doi.org/10.18653/v1/P17-2034

[243] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *ACL*. ACL, Florence, Italy, 6418–6428. https://doi.org/10.18653/v1/P19-1644

[244] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *TACL* 7 (apr 2019), 217–231. https://doi.org/10.1162/tacl_a_00264

[245] Simon Suster and Walter Daelemans. 2018. CliCR: A Dataset of Clinical Case Reports for Machine Reading Comprehension. In *NAACL-HLT*. 1551–1563. https://doi.org/10.18653/v1/N18-1140

[246] Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019. QuaRel: A Dataset and Models for Answering Questions about Qualitative Relationships. In *AAAI 2019*.

[247] Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An Open-Domain Dataset of Qualitative Relationship Questions. In *EMNLP-IJCNLP*. ACL, Hong Kong, China, 5941–5946. https://doi.org/10.18653/v1/D19-1608

[248] Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *NAACL-HLT*. ACL, New Orleans, Louisiana, 641–651. https://doi.org/10.18653/v1/N18-1059

[249] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL*. 4149–4158. https://www.aclweb.org/anthology/papers/N/N19/N19-1421/

[250] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultimodalQA: Complex Question Answering Over Text, Tables and Images. In *ICLR*. 12. https://openreview.net/pdf/f3dad930cb55abce99a229e35cc131a2db791b66.pdf

[251] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[252] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A Data Set of Information-Seeking Conversations. In *CAIR*. Tokyo, Japan, 6. https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/Thomas-etal-CAIR17.pdf

[253] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the Baseline: Single Modality Performance on Visual Navigation & QA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 1977–1983. https://www.aclweb.org/anthology/papers/N/N19/N19-1197/

[254] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[255] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 32–41. https://doi.org/10.1145/3176349.3176387

[256] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Vancouver, Canada, 191–200. https://doi.org/10.18653/v1/W17-2623

[257] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An Overview of the BIOASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics* 16, 1 (apr 2015), 138. https://doi.org/10.1186/s12859-015-0564-6

[258] Bo-Hsiang Tseng, Sheng-syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine. In *Interspeech 2016*. 2731–2735. https://doi.org/10.21437/Interspeech.2016-876

[259] Shyam Upadhyay and Ming-Wei Chang. 2017. Annotating Derivations: A New Evaluation Strategy and Dataset for Algebra Word Problems. In *ACL*. ACL, Valencia, Spain, 494–504. https://www.aclweb.org/anthology/E17-1047

[260] Svitlana Vakulenko and Vadim Savenkov. 2017. TableQA: Question Answering on Tabular Data. *arXiv:1705.06504 [cs]* (aug 2017). arXiv:cs/1705.06504 http://arxiv.org/abs/1705.06504

[261] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best Practices for the Human Evaluation of Automatically Generated Text. In *Proceedings of the 12th International Conference on Natural Language Generation*. ACL, Tokyo, Japan, 355–368. https://doi.org/10.18653/v1/W19-8643

[262] Elke van der Meer, Reinhard Beyer, Bertram Heinze, and Isolde Badel. 2002. Temporal Order Relations in Language Comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 28, 4 (jul 2002), 770–779.

[263] Teun A. van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press, New York.

[264] David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. *arXiv:1906.04701 [cs]* (jun 2019). arXiv:cs/1906.04701 http://arxiv.org/abs/1906.04701

[265] Ellen M. Voorhees and Dawn M. Tice. 2000. Building a Question Answering Test Collection. In *SIGIR (SIGIR '00)*. ACM, New York, NY, USA, 200–207. https://doi.org/10.1145/345508.345577

[266] Eric Wallace and Jordan Boyd-Graber. 2018. Trick Me If You Can: Adversarial Writing of Trivia Challenge Questions. In *ACL-SRW*. ACL, 127–133. http://aclweb.org/anthology/P18-3018

[267] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. *EMNLP* (2019). arXiv:1908.07125 http://arxiv.org/abs/1908.07125

[268] Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020. ReCO: A Large Scale Chinese Reading Comprehension Dataset on Opinion. In *AAAI*. 8. https://www.aaai.org/Papers/AAAI/2020GB/AAAI-WangB.2547.pdf

[269] Jiexin Wang, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2021. Improving Question Answering for Event-Focused Questions in Temporal Collections of News Articles. *Information Retrieval Journal* 24, 1 (feb 2021), 29–54. https://doi.org/10.1007/s10791-020-09387-9

[270] Takuto Watarai and Masatoshi Tsuchiya. 2020. Developing Dataset of Japanese Slot Filling Quizzes Designed for Evaluation of Machine Reading Comprehension. In *LREC*. ELRA, Marseille, France, 6895–6901. https://www.aclweb.org/anthology/2020.lrec-1.852

[271] Dirk Weissenborn, Pasquale Minervini, Isabelle Augenstein, Johannes Welbl, Tim Rocktäschel, Matko Bošnjak, Jeff Mitchell, Thomas Demeester, Tim Dettmers, Pontus Stenetorp, and Sebastian Riedel. 2018. Jack the Reader – A Machine Reading Framework. In *ACL*. ACL, Melbourne, Australia, 25–30. https://doi.org/10.18653/v1/P18-4005

[272] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv preprint arXiv:1502.05698* (2015).

[273] Michael White, Graham Chapman, John Cleese, Eric Idle, Terry Gilliam, Terry Jones, Michael Palin, John Goldstone, Mark Forstater, Connie Booth, Carol Cleveland, Neil Innes, Bee Duffell, John Young, Rita Davies, Avril Stewart, Sally Kinghorn, Terry Bedford, Monty Python (Comedy troupe), Python (Monty) Pictures, and Columbia TriStar Home Entertainment (Firm). 2001. Monty Python and the Holy Grail.

[274] Yuk Wah Wong and Raymond Mooney. 2006. Learning for Semantic Parsing with Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. ACL, New York City, USA, 439–446. https://www.aclweb.org/anthology/N06-1056

[275] Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In *ACL*. ACL, Florence, Italy, 5020–5031. https://doi.org/10.18653/v1/P19-1496

[276] Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. 2020. MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization. In *ACL*. ACL, Online, 3586–3596. https://www.aclweb.org/anthology/2020.acl-main.330

[277] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *EMNLP*. Lisbon, Portugal, 17-21 September 2015, 2013–2018. http://aclweb.org/anthology/D15-1237

[278] Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-Domain Question Answering on TV Show Transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. ACL, Stockholm, Sweden, 188–197. https://doi.org/10.18653/v1/W19-5923

[279] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In *EMNLP*. ACL, Brussels, Belgium, 2369–2380. http://aclweb.org/anthology/D18-1259

[280] Mark Yatskar. 2019. A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *NAACL-HLT*. 2318–2323. https://www.aclweb.org/anthology/papers/N/N19/N19-1241/

[281] Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2021. On the Faithfulness Measurements for Model Interpretations. *arXiv:2104.08782 [cs]* (apr 2021). arXiv:cs/2104.08782 http://arxiv.org/abs/2104.08782

[282] Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020. Towards Data Distillation for End-to-End Spoken Conversational Question Answering. *arXiv:2010.08923 [cs, eess]* (oct 2020). arXiv:cs, eess/2010.08923 http://arxiv.org/abs/2010.08923

[283] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2019. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJgJtT4tvB

[284] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *EMNLP*. ACL, Brussels, Belgium, 93–104. http://aclweb.org/anthology/D18-1009

[285] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In *ACL 2019*. arXiv:1905.07830 http://arxiv.org/abs/1905.07830

[286] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv:1810.12885 [cs]* (oct 2018). arXiv:cs/1810.12885 http://arxiv.org/abs/1810.12885

[287] Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020. When Do You Need Billions of Words of Pretraining Data? *arXiv:2011.04946 [cs]* (nov 2020). arXiv:cs/2011.04946 http://arxiv.org/abs/2011.04946

[288] Zhuosheng Zhang and Hai Zhao. 2018. One-Shot Learning for Question-Answering in Gaokao History Challenge. In *Proceedings of the 27th International Conference on Computational Linguistics*. ACL, Santa Fe, New Mexico, USA, 449–461. https://www.aclweb.org/anthology/C18-1038

[289] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *ICML*. arXiv:2102.09690 http://arxiv.org/abs/2102.09690

[290] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language Using Reinforcement Learning. *arXiv:1709.00103 [cs]* (nov 2017). arXiv:cs/1709.00103 http://arxiv.org/abs/1709.00103

[291] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a Vacation" Takes Longer than "Going for a Walk": A Study of Temporal Commonsense Understanding. In *EMNLP-IJCNLP*. ACL, Hong Kong, China, 3361–3367. https://doi.org/10.18653/v1/D19-1332

[292] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *ACL*. https://arxiv.org/abs/2105.07624

[293] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-Domain Question Answering. *arXiv:2101.00774 [cs]* (jan 2021). arXiv:cs/2101.00774 http://arxiv.org/abs/2101.00774

[294] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. 2017. Uncovering the Temporal Context for Video Question Answering. *International Journal of Computer Vision* 124, 3 (sep 2017), 409–421. https://doi.org/10.1007/s11263-017-1033-7

[295] Rolf A. Zwaan. 2016. Situation Models, Mental Simulations, and Abstract Concepts in Discourse Comprehension. *Psychonomic Bulletin & Review* 23, 4 (aug 2016), 1028–1034. https://doi.org/10.3758/s13423-015-0864-x