Unsupervised Question Answering: Challenges, Trends, and Outlook

Pratyay Banerjee **Tejas Gokhale** **Chitta Baral**
Arizona State University
{pbanerj6, tgokhale, chitta}@asu.edu

Abstract

Question answering (QA) is considered to be a central aspect of natural language processing (NLP) and has seen remarkable progress in the last decade, brought-about by transformer-based language models trained on large human-annotated text corpora. However, several pitfalls of supervised training have been identified, especially when considering performance of such systems on new domains, linguistic styles, and adversarial samples. Unsupervised question answering – the ability to answer questions without explicit supervision from human-annotated training data, has emerged as a research direcftion that could potentially mitigate these pitfalls. This paper reviews recent trends in unsupervised question answering and provides a unifying perspective of work in this area, along with a survey of the closely related directions of weakly and partially supervised QA models. We provide insights into associated challenges and potential research directions towards robust unsupervised QA models.

1 Introduction

Question-answering (QA) is considered to be integral to the human reasoning process (Turing, 1950) and the development of systems that resemble this ability has been a long-standing research program in natural language processing (Simmons, 1965). QA systems are crucial for evaluating natural language understanding and human-machine communication via dialog systems. Several datasets have been proposed for QA tasks such as extractive question answering (predicting a span of text as answer) (Rajpurkar et al., 2018; Yang et al., 2018; Kwiatkowski et al., 2019) and multiple-choice question answering (predicting an answer from a list of choices) (Sap et al., 2019; Talmor et al., 2019; Zellers et al., 2018; Clark et al., 2018). Many of these tasks require reasoning over contexts, corpora, and commonsense and scientific knowledge.

Large pre-trained language models (PLMs) (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Brown et al., 2020) have resulted in significant performance improvements on these tasks, using fully-supervised training protocols. Unfortunately, these methods overfit the training data and do not transfer well to new domains, especially for low-resource domains where large-scale training data collection may not be feasible. Spurious correlations, annotation artifacts, and linguistic biases in NLP datasets also affect generalization (Gururangan et al., 2018; Niven and Kao, 2019; Kaushik and Lipton, 2018; Poliak et al., 2018). Analysis of BERT embeddings reveals artifacts such as two random words having high cosine similarity (Ethayarajh, 2019), and 25% tokens being assigned to incorrect clusters (Mickus et al., 2019). PLMs also fail in question-answering tasks with negated questions in cloze completion (Kassner and Schütze, 2020; Ettinger, 2020), multiple-choice QA (Asai and Hajishirzi, 2020), and visual question answering (Gokhale et al., 2020). These findings are undesirable for robustness considerations. While carefully-designed crowd-sourcing (Sakaguchi et al., 2020) and dataset filtering (Le Bras et al., 2020) have been suggested to mitigate these phenomena, these are typically associated with a high cost of data annotation.

This survey focuses on various efforts towards unsupervised question answering (on English language inputs) . While task-specific (Wang, 2006; Wu et al., 2017; Fu et al., 2020; Zhu et al., 2021) and method-specific (Lai et al., 2018; Storks et al., 2019) surveys of question answering and review of recent datasets (Rogers and Rumshisky, 2020) are available, this paper is the first survey on unsupervised QA, drafted with the following objectives:

1. to review recent development of QA models trained without explicit supervision,
2. to identify key challenges in unsupervised QA,
3. to recommend potential research directions to mitigate these challenges.

The paper is structured as follows. Section 2 introduces the problem setup for unsupervised question answering, and provides a categorization of various QA tasks and major evaluation benchmarks. Section 3 surveys existing methodologies, training protocols, and results for unsupervised QA models. Section 4 discusses the related problems of learning from weak and partial supervision. Finally, we delineate challenges associated with unsupervised methods in Section 5, and offer our insights in Section 6 to open up potential research directions for future work in this area.

2 Unsupervised Question Answering

Problem Setup: In the unsupervised question answering setup, typically, a dataset of context paragraphs is available, and the model must learn to answer questions about these paragraphs. In some cases, a set of questions may also be provided as part of the dataset; however the true answers to each question are not available during training.

We consider four categories under this problem setup, for which unsupervised QA methods have been explored: Winograd Schema Challenge (WSC), Extractive QA (EQA), Multiple-Choice QA (MCQA), and Multi-Modal QA. We distinguish WSC as a separate category as it only has a test set which necessitates unsupervised or commonsense knowledge acquisition methods, and could be treated as either a classification, extractive, or a generative task. Furthermore, it has been studied as an unsupervised problem for several years.

2.1 Winograd Schema Challenge

Inspired by examples from Winograd (1972) illustrating the challenges of natural language understanding and the importance of contextual knowledge, the Winograd Schema Challenge (WSC) was proposed by Levesque et al. (2012) and further developed by Morgenstern et al. (2016). An example from WSC is shown below:

WSC item: The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.
Question: Who [feared/advocated] violence?

Winograd Schemas (sentences and questions containing pronouns), are provided as input, and the system must resolve the entity that the pronoun refers to. If the co-referent is changed from *feared* to *advocated* in both the sentence as well as the question, the answer changes from *councilmen* to *demonstrators*. The WSC challenge does not provide a training dataset, but only a test set for evaluating systems – this set originally had 60 samples which have now grown to 273 or 285. As such, there is no explicit supervision available to train machine learning models.

However, large QA datasets for pronoun resolution have been compiled, such as the Definite Pronoun Resolution Dataset (Rahman and Ng, 2012), Winogender (Rudinger et al., 2018) where the pair of sentences differ only by gender, and KnowRef (Emami et al., 2019) with ambiguous pronominal anaphora, and the WinoGrande (WG) (Sakaguchi et al., 2020) which is a crowdsourced dataset of $44k$ samples with training-development-test splits. Table 1 suggests that the supervised RoBERTa model, trained on the WG corpus is able to achieve a high accuracy of 90.1% on the WSC test set. However, the same model results in a lower accuracy of 79.4% on the WG test set. Sakaguchi et al. (2020) have postulated that the model might be picking up spurious correlations in WSC, while at the same time being unable to generalize on the WG test set itself. Thus we argue that WSC and WSC-style challenges are far from solved, motivating research into unsupervised methods in this domain to address the issue of spurious correlations and linguistic biases.

2.2 Extractive QA (EQA)

Extractive QA or Reading Comprehension, is the task in which a text "context" or passage is provided as input along with a question, and EQA systems are expected to extract the answer as a span of text in the context. Multiple datasets have been developed for EQA that we describe below.

SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016) contains $100k$ open-ended questions based on context passages from Wikipedia articles. Answers to these questions are present explicitly in the context and do not require commonsense reasoning over the context. Following is an example:

Paragraph: In February 2016, over a hundred thousand people signed a petition in just twenty-four hours, calling for a boycott of Sony Music and all other Sony-affiliated businesses after rape allegations against music producer Dr. Luke were made by musical artist Kesha. Kesha asked a New York City Court to free her from her contract with Sony, but the court denied the request.
Question: How many people signed a petition to boycott Sony Music in 2016?
Answer: over a hundred thousand

SQuAD 2.0 (Rajpurkar et al., 2018) was proposed as an addendum to SQuAD. It contains a set of $50k$ "unanswerable" questions, i.e. questions that do not have answers explicitly in the provided context but may require systems to use external knowledge and reasoning to find the answer.

NewsQA (Trischler et al., 2017) contains over $100k$ Q-A pairs crowd-sourced from $10k$ CNN news articles (Hermann et al., 2015), with answers being text-spans in the articles. The dataset was curated such that question-answering would require reasoning skills. Subsequently, datasets for advanced reasoning tasks have been proposed, such as **HotPotQA** (Yang et al., 2018) which requires multi-hop reasoning, and **Natural Questions** (Kwiatkowski et al., 2019) which contains questions entered into search engines by real users. The data collection protocol for NQ, where the users actively search for unknown answers to their questions, is markedly different from previous work where the question annotators typically know the answer to their own question (Lee et al., 2019).

2.3 Multiple-choice QA (MCQA)

In contrast to extractive QA, in a multiple-choice question answering (MCQA) task, a list of answer choices is provided as input. Thus the system must interpret the question and predict an answer from one of these choices. Datasets developed for MCQA are listed below.

CommonsenseQA (Talmor et al., 2019) is a five-way multiple-choice QA benchmark containing 9500 questions. Each question requires disambiguation of a target concept from three connected concepts. These connected concepts come from *ConceptNet* (Liu and Singh, 2004), which is a large knowledge-base that capture a diverse range of commonsense concepts and relations about spatial, physical, social, temporal, and psychological aspects of everyday life. As such, a QA task constructed using ConceptNet is challenging.

aNLI (Bhagavatula et al., 2019) is intended to judge the abductive reasoning ability of QA systems to form possible explanations for a given set of observations. The task is to find a hypothesis (from a list of choices) that explains an input "post-observation" given a "pre-observation". As such, the task calls for an understanding of the sequential occurrence of events. Following is an example:

Observation 1: Jim was working on a project.
Observation 2: Luckily, he found it in a nearby shelf.
Hypothesis 1: Jim found he was missing an item. ✓
Hypothesis 2: Jim needed a certain animal for it. ✗

SocialIQA (Sap et al., 2019) is a dataset containing 3-way multiple-choice questions that require reasoning about social interactions and implications of events, given a passage about a social situation as context. Several question types in this dataset are derived from the *Atomic* inference dimensions (Sap et al., 2019), such as actor *intention*, actor *motivation*, *effect* on the actor and others, etc.

Science-based Question Answering: Several MCQA datasets require an ability to answer scientific questions at different difficulty levels. The AI2 Reasoning Challenge (ARC) (Clark et al., 2018) contains 8000 four-way multiple-choice science questions and answers along with a large corpus of 14 million scientific facts that are necessary to answer the questions. These questions require multi-hop reasoning, i.e. the ability to combine information spread over multiple disconnected facts. OpenBookQA (Mihaylov et al., 2018) is a 4-way MCQA dataset, for which partial information from a small corpus of 3000 facts is necessary to answer the question. Systems are free to retrieve the other partial information from any external source. QASC (Khot et al., 2020), is an 8-way MCQA dataset, for which questions can be answered by exactly two facts from an associated corpus.

2.4 Multi-modal QA

Question-answering has also been extended to questions about images or videos. VQA-v2 (Goyal et al., 2017), VizWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019), and CLEVR (Johnson et al., 2017) are major benchmarks for image-based question answering, where the answers are open-ended words or short phrases. VQA-CP-v2 (Agrawal et al., 2018) is a reorganization of VQA-v2 that seeks to measures the out-of-distribution generalization ability of the question answering system. Reasoning aspects have also been explored for multi-modal QA, such as Visual Commonsense Reasoning (Zellers et al., 2019) focusing on commonsense reasoning and rationalizing in a four-way multiple-choice task, OK-VQA (Marino et al., 2019) that requires reasoning with external knowledge, VQA-LOL (Gokhale et al., 2020) focusing on logical questions, and introspective sub-questions in (Selvaraju et al., 2020). In the domain of video question answering, VideoQA (Yang et al., 2003), MSR-VTT-QA (Xu et al., 2017), MovieQA (Tapaswi et al., 2016), and TVQA (Lei et al., 2018) have been proposed.

Approach	Accuracy
RoBERTa-WG (Sakaguchi et al., 2020)*	**90.1**
K-Parser (Sharma et al., 2015)	53.0
Modified Skip-Gram (Zhang and Song, 2018)	60.3
BERT Inner Attention (Klein and Nabi, 2019)	60.3
BERT-MASKEDWIKI (Kocijan et al., 2019)	61.9
UDSSM (Wang et al., 2019)	62.4
Ensemble LMs (Trinh and Le, 2018)	63.7
CSS (Klein and Nabi, 2020)	69.6
GPT-2 (Brown et al., 2020)	70.7
WSC Knowledge Hunting (Prakash et al., 2019)	71.1

Table 1: Comparison of the different unsupervised methods on the Winograd Schema Challenge. (*) indicates supervised method.

3 Unsupervised Methods for QA

In this section, we describe the different approaches to unsupervised QA. Results on the respective benchmark datasets are shown in Tables 1, 2, and 3.

3.1 Winograd Schema Challenge

Semantic Parsing and Sample-guided Graph-based Reasoning. The method in (Sharma et al., 2015) utilizes semantic parsing and information retrieval to gather similar sentences with disambiguated pronouns using the original schema sentence as a query. Question answering is guided using a graph-based reasoning algorithm defined over the output of the semantic parser, exploiting the retrieved unambiguous sentence structure.

Skip-Gram and Semantic Dependencies Pre-Training. Zhang and Song (2018) propose a modified skip-gram objective for pre-training word embeddings to predict semantic dependencies between verbs. A set of vector-space models are trained to capture the verb meaning and transferred to related ambiguous pronouns.

Word Attention Scores. Wang et al. (2019) propose Unsupervised Deep Structured Semantic Models (UDSSM), in which a BiLSTM is trained to compute contextual word embeddings and use the word attention scores between ambiguous pronouns and the noun as the prediction scores. Extending the previous work, Klein and Nabi (2019) directly exploit the inner attention layers of BERT to compute a maximum over the attention scores between the pronoun and the noun.

Pre-training on Masked Noun or Entity Prediction. Kocijan et al. (2019) construct a synthetic dataset called MaskedWiki, crawled from English Wikipedia to pre-train a language model for a synthetic masked-noun prediction pseudo-task. In this task, a noun-word is masked, and the model is asked to predict the word. Ye et al. (2019) adopt a "align, mask, and select (AMS)" strategy where entities that are connected with ConceptNet are masked, and the model is asked to predict among a list of similar candidate entities.

Large Language Models. An ensemble of large pre-trained models was first utilized by Trinh and Le (2018) and GPT is evaluated on WSC by Brown et al. (2020). Prakash et al. (2019) extend a language model with a knowledge hunting strategy using a probabilistic soft-logic framework with hand-crafted rules and entity alignment strategy. A similar knowledge-hunting approach is evaluated on Winogrande dataset by Sakaguchi et al. (2020).

Contrastive Self-Supervision. Klein and Nabi (2020) study a self-supervised learning approach by exploiting the structural information present in Winograd Schema pairs – if one word is changed, the pronoun becomes the coreference of a different noun. A contrastive margin loss is defined to operate on a particular sentence's probable answer candidates and a mutual exclusion loss operating on a pair of sentences.

3.2 Extractive QA

Unlike the unsupervised methods for WSC which acquire commonsense knowledge from word embeddings, knowledge hunting, or large-scale pre-training of language models, unsupervised methods for EQA focus on synthesizing question-answer pairs given a text passage. Using these synthetic data, a QA model can be trained, and evaluated on existing human-authored EQA benchmarks described in Section 2.2. Below, we discuss various question-answer pair generation methods.

Cloze Generation. In Cloze Generation, a textual passage is divided into a preliminary introduction P and a trailing part from which the question Q and the answer A are selected. The answer-span is selected first, such that it is present in both the premise and question, and is replaced with a placeholder in the question as shown below:

> **Passage:** Autism is a neuro-developmental disorder characterized by impaired social interaction, verbal and non-verbal . . .
> **Question:** People with autism tend to be a little aloof with little to no ________.
> **Answer:** social interaction.

Cloze generation for training was proposed Dhin-

	SQuAD 1.1	NewsQA
BERT-Large (*)	85.1 / 91.8	N/A / 73.6
BERT-Large +		
(Dhingra et al., 2018)	28.4 / 35.8	18.6 / 27.2
(Lewis et al., 2019)	44.2 / 54.7	17.9 / 27.0
(Fabbri et al., 2020)	46.1 / 56.8	21.2 / 29.4
(Li et al., 2020)	61.1 / 71.4	32.1 / 45.1

Table 2: Comparison of different unsupervised methods on extractive QA task. Exact Match and F1 scores are reported. (*) indicates supervised method.

gra et al. (2018), with ground-truth answer-spans being a sequence of overlapping text between the introduction passage and the trailing part.

Unsupervised Cloze Translation. On the other hand, Lewis et al. (2019) select answer spans from noun-phrases as well as named-entities, and present four methods of unsupervised cloze translation, adapted to convert a cloze-style question-answer pair to a more natural question-answer pair: (1) Identity Mapping, where original cloze-style pairs are evaluated, (2) Clozes, where a random perturbation, word-ordering change, and random or heuristics based "Wh-word" is prepended, (3) rule-based question generation (Heilman and Smith, 2010) using Wh-movement via syntactic transformation, and (4) a Seq-2-Seq neural model trained in an unsupervised fashion with two non-parallel training corpus, the source Cloze-style questions, and the target natural questions. The training process is similar to translation models (Lample et al., 2018) with a bidirectional combination of in-domain training using denoising autoencoding and cross-domain training using online-back-translation.

Retrieval and Template-based Question Generation. Fabbri et al. (2020) propose a two-step method as an extension to the above work. First, the context is used to retrieve similarly-structured sentences. These sentences are then used to generate questions using template-based methods. Given a context of the format:

[FRAGMENT I][ANSWER][FRAGMENT II]

a template of the form: *"Wh* + II + I + *?"* is used to construct the question, with a *Wh*-word replacing the answer-word in the question.

RefQA and Iterative Refinement. There are several limitations of using Cloze Generation as the only source of question-answer pair generation. There are significant lexical overlaps between the generated questions and the paragraph, which allows the QA model to predict the answer simply via word matching, thereby affecting generalization. Moreover, the answer category is limited to the named entity or noun phrase, further restricting the model's coverage. To mitigate these challenges, Li et al. (2020) propose RefQA, which utilizes cited documents in parent Wikipedia context documents to extract clozes with minimal text overlap with parent context. Furthermore, they propose a dependency-parsing-based cloze-translation to natural questions. First, the right child nodes of the answer are retained, and the left children are pruned. Second, if the child node's subtree contains the answer for each node of the parse tree, the child node is moved to the first child node. Finally, an in-order traversal is performed over the reconstructed tree. A rule-based mapping is applied to replace the special mask token of the cloze with an appropriate "Wh-word".

In Iterative Refinement, a neural model is first trained with a generated question-answer pair. This model is used for answer prediction to generate a new answer $\hat{A}$. If $\hat{A}$ is different from the original answer A, then this new answer span is used as a seed for a new question generation $\hat{Q}$ using the above method. This process is repeated till no new Q, A pairs are generated.

Multi-hop Question Generation (Pan et al., 2020) utilizes multiple parallel data sources, such as tables and associated paragraphs. A fixed set of operators is defined to extract, generate, aggregate, or merge information. Six pre-defined reasoning graphs (similar to action templates) are used for generating multi-hop questions.

3.3 Multiple-choice QA

Unsupervised MCQA methods rely on external knowledge graphs such as *Atomic* (Sap et al., 2019) and ConceptNet (Liu and Singh, 2004), or additional factual sentences as provided in the ARC, QASC, and OpenBookQA datasets. Some methods also use large language models such as GPT-2 and Comet (Bosselut et al., 2019).

Information Retrieval Solver was proposed in ARC (Clark et al., 2016), in which *(context, question, answer)* options are used as queries. The top retrieved sentence with a non-stop-word overlap with the question-answer pair is used as a representative, and its corresponding ranking score (BM25)

	CSQA	aNLI	SIQA	ARC	QASC	OBQA
Random	20.0	50.0	33.3	25.0	12.5	25.0
RoBERTa (*)	78.5	85.6	76.6	67.0	61.8	72.0
RoBERTa	45.0	65.5	47.3	23.8	23.8	19.7
GPT-2	41.4	56.5	44.6	25.0	13.2	27.0
IR Solver	24.4	54.8	36.0	21.2	19.4	28.8
Self-Talk	32.4	N/A	46.2	N/A	N/A	N/A
Dynamic Gr.	N/A	N/A	50.1	N/A	N/A	N/A
Know. Trip. L.	38.8	65.3	48.5	28.4	27.2	33.8
Dataset Cons.	67.9	70.8	63.2	N/A	N/A	N/A

Table 3: Comparison of classification accuracies for different unsupervised methods on multiple-choice QA task. (*) indicates supervised method.

is used as answer confidence. The option with the highest score is chosen as the answer.

Self-Talk (Shwartz et al., 2020) is an unsupervised framework inspired by inquiry-based discovery learning. In this approach, the system inquires a language model such as GPT-2 or Comet with several information-seeking questions such as *"what is the definition of [concept]"* to discover additional background knowledge. After an answer is generated, the method utlizes these additional question-answer pairs as context. Finally, the answer is selected from the given choices using the least cross-entropy score for the sequence of text generated by concatenating the generated context, question, and the answer option.

Self-Supervised Knowledge Triplet Learning (Banerjee and Baral, 2020) was proposed to pretrain large language models such as RoBERTa, with three representation learning functions that aim to complete a knowledge triple given two of its elements. For example, given a *(context, question, answer)* triple, one function generates the context given the QA pair, another generates the question given the context and the answer. These functions are used in conjunction to compute the distance for each answer candidate from the generated answer representation. Methods for knowledge graph construction from unstructured text corpora are proposed that use noun/verb phrases to create knowledge triples required for pre-training.

Dynamic Neuro-Symbolic Knowledge Graph Construction. In Bosselut et al. (2021), an initial study on zero-shot commonsense question answering is conducted by formulating the task as inference over dynamically generated commonsense knowledge graphs. In contrast to prior studies for knowledge integration that rely on retrieval from static knowledge graphs, this work requires commonsense knowledge integration where contextually relevant knowledge is often not present in existing knowledge bases. The method generates contextually-relevant symbolic knowledge structures "on-demand" using generative neural commonsense knowledge models such as Comet and GPT-2. The method defines a reasoning algorithm using this "on-demand" generated knowledge graphs and selects the most supported answer option from the additional knowledge context.

Knowledge-driven Data Construction. In Ma et al. (2021), a neuro-symbolic framework for zero-shot question answering across commonsense tasks is proposed. Guided by a set of hypotheses, the framework studies how to transform various pre-existing knowledge resources into a most effective form for pretraining models. The framework varies the set of language models, training regimes, knowledge sources, and data generation methods and measures their impact across tasks. Extending on Self-Talk and Knowledge Triplet Learning, it compares and contrasts four constrained distractor-sampling strategies. The key insight derived from the work is while an individual knowledge graph is better suited for specific tasks, a global knowledge graph brings consistent gains across different tasks. Also, preserving the task structure and generating questions that are fair and informative helps large language models learn more effectively.

3.4 Multi-Modal Question Answering

There are few unsupervised methods for VQA and video-QA where human-authored QA pairs are unavailable. We categorize the methods in two categories, the first being unsupervised methods for *out-of-vocabulary generalization*, and the second being *weakly supervised QA* in which no human-authored QA pairs are available, but other signals such as captions and transcriptions can be used.

Zero-Shot VQA. In this task, the systems are expected to generalize to out-of-vocabulary questions or answers during test-time. The task was first proposed in (Teney and Hengel, 2016), in which they introduced multiple methods based on pre-trained word embeddings, object classifiers with semantic embeddings, and test-time retrieval of example images that are encoded in a semantic embedding space. The final answer is generated using a look-up table and nearest neighbor search in answer-embedding space.

Unsupervised Task Discovery proposed by Noh et al. (2019), utilizes existing large-scale visual datasets with annotations such as image class labels, bounding boxes, and region descriptions to learn rich and diverse visual concepts. The missing link between question-dependent answering models and *visual data without questions* makes learning visual concepts challenging. This is mitigated by learning a task conditional visual classifier capable of solving diverse question-specific visual recognition tasks, and transfering the classifier to VQA models. To learn the unsupervised task discovery, external structured knowledge sources such as ConceptNet and WordNet are utilized.

Weakly Supervised from Captions Two recent papers utilize captions to generate QA pairs for image-based VQA and video QA, respectively. Both the methods have shown a competitive performance to existing supervised methods.

Banerjee et al. (2020) utilize various question generation techniques such as cloze-generation, template-based methods, and semantic role-labeling, using the image captions as context. Paraphrasing using back-translation is employed for linguistic diversity. Particular object entity-based and yes/no based questions are generated following the process introduced in COCO-QA (Ren et al., 2015). In semantic role labeling (FitzGerald et al., 2018), the role-labels are expressed as question-answer pairs. For example, for the caption *"A girl in a red shirt holding a skateboard sitting in an empty open field"*, Q-A pairs such as *("What is someone holding?", "a skateboard")* are generated.

In (Yang et al., 2020), captions for a huge set of videos are generated using automated speech recognition. A pre-trained transformer model on SQuAD is used for generating question-answer pairs from these pre-processed captions.

4 Related Paradigms of Learning

Self-Supervised Pre-Training. Self-supervised learning leverages auxiliary tasks with input-output samples extracted from unlabeled datasets, to learn generalizable representations applicable to multiple downstream tasks. Self-supervision has been used to train transformer-based language models using masked token prediction (Devlin et al., 2019; Raffel et al., 2020), sequence prediction (Yang et al., 2019), discriminator-based plausible alternative prediction (Clark et al., 2019).

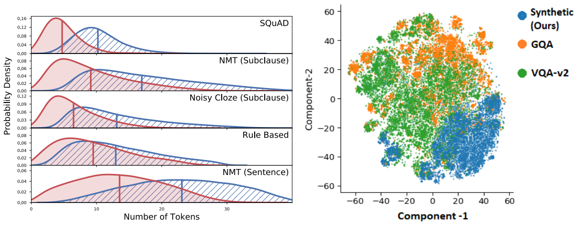

Figure 1: Discrepancy between dataset questions and generated questions. *Left:* Plot from Lewis et al. (2019) showing a comparison of question lengths for various generation methods. *Right:* tSNE plot from Banerjee et al. (2020) comparing question embeddings for VQA.

Low-Resource Question Answering. In many cases, training datasets for QA may be small in size, thereby affecting model generalization. To alleviate this, methods utilizing reinforcement learning for question generation (Yang et al., 2017), cloze question generation (Dhingra et al., 2018), and meta learning (Yan et al., 2020) have been proposed.

Zero-Shot and Few-Shot Learning. An approach is to utilize domain adaptation methods to train the model on a large-scale source task and to finetune it on the low-resource target task (Kadlec et al., 2016; Golub et al., 2017; Wiese et al., 2017; Chung et al., 2018). However this approach assumes access to a labeled source dataset. Recently, GPT-3 (Brown et al., 2020), a large language model (175B parameters) has been trained with huge text corpora (300B tokens). While GPT-3 is able to perform a wide variety of NLP tasks after this expensive pre-training, the zero-shot performance is still below some unsupervised methods discussed in this survey, such as 70.2% on WSC and 59.5% on SQuAD-v2. This in our opinion, makes a strong case for further research in unsupervised learning, especially with regards to generalization.

5 Challenges

Aforementioned methods for unsupervised QA have unveiled challenges related to reasoning abilities and generalization that need to be addressed. We discuss these challenges below.

Question-Answer Pair Generation. Although question-answer pair generation has improved a lot over the years, there is still a gap to fill that is observed when purely unsupervised methods are compared to self-training methods such as (Alberti et al., 2019; Puri et al., 2020) that use human-authored questions and answers to train question-generation models and then train neural readers only using the generated synthetic question-answer pairs. Figure 1 shows the gap between generated questions (Lewis et al., 2019) and original SQuAD dataset distribution (left), and VQA-v2 and GQA vs. synthetic questions from (Banerjee et al., 2020). Further improving non-parallel unsupervised cloze translation, utilizing existing lexical and knowledge graphs for additional supervision, and improving parsing-based question generation would be an interesting direction to bridge this gap.

Answer-Phrase Generation. Named-entities and noun-phrases are the current focus for answer generation. While recent methods (Banerjee et al., 2020) have introduced semantic-role labeling to generate a answer-phrases with diversity in parts-of-speech generated, there remains a large room for improving synthetic answer generation.

Training Sample Selection. As the procedural question-answer pair generation does not restrict the size of the synthetic training corpus, there is a limit to positive inductive bias that can be incorporated into certain neural architectures, limiting the generalization ability and moving towards overfitting to the synthetic corpus. Utilizing train sample selection, adversarial sample selection, hard-sample mining, and curriculum learning would be the next step to understand which samples are more useful to learn question answering.

Reasoning Abilities. Although commonsense reasoning is required in WSC, aNLI, and other commonsense-related tasks, other tasks such as complex multi-hop reasoning, abductive reasoning where the hypotheses are generated and not selected, quantitative, temporal, qualitative, and non-monotonic reasoning, all remain uphill battles. Similarly, in visual question answering, unsupervised question-answer pair generation with complex spatial reasoning in focus is still unexplored. Meanwhile (Ye and Kovashka, 2021) have shown that supervised models can take advantage of shortcuts and co-occurring words between the question and answer-choices in VCR (Zellers et al., 2019). Unsupervised learning could help break these spurious shortcuts in order to boost generalization.

Evaluation Metrics used in current question answering benchmarks range from classification accuracy for multiple-choice QA, exact match, and $F1$-score for extractive QA, to a custom visual question answering metric incorporating multiple allowed phrases for VQA tasks. While there has been work towards generative question answering models (Bhakthavatsalam et al., 2021), existing evaluation metrics designed for classification or MCQA tend to over-penalize methods that generate correct but descriptive answers (Goyal et al., 2017; Banerjee et al., 2020). It is intractable to annotate datasets with all possible answers to a question given that some questions may be subjective and have multiple answers, and in lieu of the plethora of synonymous or equivalent phrases in natural language. Hence, there is a need for newer metrics that judge multi-word descriptive paraphrased versions of the correct answer equally. While the issue of better evaluation has attracted attention for the tasks of machine translation (Edunov et al., 2020) and text generation systems (Gehrmann et al., 2021), it remains under-explored in the QA domain, with few works such as (Luo et al., 2021) which seeks to develop automated methods to augment answer annotations with equivalent and alternate answers.

6 Outlook

In a typical QA setting, specific words in the text may not be enough to answer the question since contextual knowledge may be required, as is aptly highlighted by the Winograd Schema Challenge. Collection of such external knowledge covering a wide range of knowledge and reasoning abilities is often infeasible. Therefore development of techniques that do not rely on the collection of datasets is important for low-resource settings and for adapting models to new domains, or when the knowledge-base changes over time – for instance Wikipedia entries on most topics are updated over time. There has been recent interest in "Test-Time Training" (Sun et al., 2020) for image classification –an approach that turns a single unlabeled test sample into a self-supervised learning problem on which the model is trained before making a prediction. This paradigm could be potentially extended to QA tasks for improving generalization without reliance on human-authored data. Spurious correlations and biases bring in imminent risks, especially when it comes to socio-cultural biases that have been shown to percolate into training datasets. Unsupervised learning can potentially serve as a tool to not only mitigate these risks but also study their impact, as any observed biases could be attributed back to data synthesis methods.

Ethical Consideration.

Linguistic biases, social-cultural and historical biases have been shown to not only exist in human-annotated datasets used for many NLP tasks (Hendricks et al., 2018; Bender et al., 2021; Kurita et al., 2019; Sheng et al., 2019), but also result in the model learning these biases to make predictions. Unsupervised learning methods discussed above use a fixed set of question-answer pair generation. While this results in lesser diversity than natural and human-annotated corpora, it enables us to control the training data. This control over training data may allow researchers to trace back the cause of biased predictions to one or more data generation mechanisms. This could help the community identify potential sources of bias and incorporate these findings for constructing new datasets and also for debiasing existing datasets.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *ACL*.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–162.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2020. Self-supervised vqa: Answering visual questions using images and captions. *arXiv:2012.02356*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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 *ICLR*.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer ai2 reasoning challenge. *arXiv preprint arXiv:2102.03315*.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *AAAI*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Yu-An Chung, Hung-Yi Lee, and James Glass. 2018. Supervised and unsupervised transfer learning for question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1585–1594.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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*.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *NAACL-HLT*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systemstrained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *ACL*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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)*, pages 55–65.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *ACL*.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *ACL*.

Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, and Jian Sun. 2020. A survey on complex question answering over knowledge base: Recent advances and challenges. *arXiv:2007.13069*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *ECCV*.

David Golub, Po-Sen Huang, Xiaodong He, and Li Deng. 2017. Two-stage synthesis networks for transfer learning in machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 835–844.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *NAACL-HLT*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Karl Moritz Hermann, Tomás Kociskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

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*, pages 2901–2910.

Rudolf Kadlec, Ondřej Bajgar, Peter Hrincar, and Jan Kleindienst. 2016. Finding a jack-of-all-trades: An examination of semi-supervised learning in reading comprehension.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *EMNLP*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *AAAI*.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *ACL*.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *ACL*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the Winograd schema challenge. In *ACL*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Tom Kwiatkowski, Jennimaria Palomaki, et al. 2019. Natural questions: a benchmark for question answering research. *TACL*.

Tuan Lai, Trung Bui, and Sheng Li. 2018. A review on deep learning techniques applied to answer selection. In *COLING*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *EMNLP*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *ICML*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *EMNLP*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *ICPKRR*. Citeseer.

P Lewis, L Denoyer, and S Riedel. 2019. Unsupervised question answering by cloze translation. In *ACL*.

Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. Harvesting and refining question-answer pairs for unsupervised QA. In *ACL*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*.

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*.

Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 2021. 'just because you are right, doesn't mean I am wrong': Overcoming a bottleneck in development and evaluation of open-ended VQA tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2766–2771, Online. Association for Computational Linguistics.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *AAAI*.

Kenneth Marino, Mohammed Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2019. What do you mean, bert? assessing bert as a distributional semantics model. *arXiv preprint arXiv:1911.05758*.

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*.

Leora Morgenstern, Ernest Davis, and Charles L Ortiz. 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *ACL*.

Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, and Bohyung Han. 2019. Transfer learning via unsupervised task discovery for visual question answering. In *CVPR*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2020. Unsupervised multi-hop question answering by question generation. *arXiv:2010.12623*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *CoNLL*.

Ashok Prakash, Arpit Sharma, Arindam Mitra, and Chitta Baral. 2019. Combining knowledge hunting and neural language models to solve the winograd schema challenge. In *ACL*.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *EMNLP*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *NIPS*.

Anna Rogers and Anna Rumshisky. 2020. A guide to the dataset explosion in qa, nli, and commonsense reasoning. In *COLING: Tutorial Abstracts*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.

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

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at vqa models: Introspecting vqa models with sub-questions. In *CVPR*.

Arpit Sharma, Nguyen Ha Vo, Somak Aditya, and Chitta Baral. 2015. Towards addressing the winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In *IJCAI*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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)*, pages 3398–3403.

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

R. F. Simmons. 1965. Answering english questions by computer: A survey. *Commun. ACM*.

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

Y Sun, X Wang, et al. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*. PMLR.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *CVPR*.

Damien Teney and Anton van den Hengel. 2016. Zero-shot visual question answering. *arXiv:1611.05546*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv:1806.02847*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *RepLNLP*.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433.

Mengqiu Wang. 2006. A survey of answer extraction techniques in factoid question answering. *Computational Linguistics*, 1(1).

Shuohang Wang, Sheng Zhang, Yelong Shen, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Jing Jiang. 2019. Unsupervised deep structured semantic models for commonsense reasoning. In *NAACL*.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural question answering at bioasq 5b. In *BioNLP 2017*, pages 76–79.

T. Winograd. 1972. Understanding natural language. *Cognitive psychology*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *CVIU*, 163.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM-Multimedia*.

Ming Yan, Hao Zhang, Di Jin, and Joey Tianyi Zhou. 2020. Multi-source meta transfer for low resource multiple-choice question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7331–7341.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2020. Just ask: Learning to answer questions from millions of narrated videos. *arXiv:2012.00451*.

Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. 2003. Videoqa: question answering on news video. In *ACM-Multimedia*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NIPS*.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised qa with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050.

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*.

K Ye and A Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *AAAI*.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv:1908.06725*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Hongming Zhang and Yangqiu Song. 2018. A distributed solution for winograd schema challenge. In *2018 10th ICML and Computing*.

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*.