Natural Language QA Approaches using Reasoning with External Knowledge

Chitta Baral[1*] , **Pratyay Banerjee**[1] , **Kuntal Pal**[1] and **Arindam Mitra**[2]

[1]Arizona State University
[2]Microsoft Inc.
{chitta,pbanerj6,kkpal}@asu.edu, arindam.mitra@microsoft.com

Abstract

Question answering (QA) in natural language (NL) has been an important aspect of AI from its early days. Winograd's "councilmen" example in his 1972 paper and McCarthy's Mr. Hug example of 1976 highlight the role of external knowledge in NL understanding. While Machine Learning has been the go-to approach in NL processing as well as NL question answering (NLQA) for the last 30 years, recently there has been an increasingly emphasized thread on NLQA where external knowledge plays an important role. The challenges inspired by Winograd's councilmen example, and recent developments such as the Rebooting AI book, various NLQA datasets, research on knowledge acquisition in the NLQA context, and their use in various NLQA models have brought the issue of NLQA using "reasoning" with external knowledge to the forefront. In this paper we present a survey of the recent work on them. We believe our survey will help establish a bridge between multiple fields of AI, especially between (a) the traditional fields of knowledge representation and reasoning and (b) the field of NL understanding and NLQA.

1 Introduction and Motivation

Understanding text often requires knowledge beyond what is explicitly stated in the text. While this was mentioned in early AI works [44], with recent success in traditional NLP tasks, many challenge datasets have been recently proposed that focus on hard NLU tasks. The Winograd challenge was proposed in 2012, bAbI [79] , and an array of datasets were proposed in 2018-19. This includes[1] QASC [34], CoQA [61] , DROP [20] , BoolQ [14] , CODAH [11] , ComQA [1] , CosmosQA [31] , NaturalQuestions [37] , PhysicalIQA [7] , QuaRTz [68] , QuoRef [18] , SocialIQA [65] , Wino-Grande [63] in 2019 and ARC [16] , CommonsenseQA [72] , ComplexWebQuestions [71] , HotpotQA [83] , OBQA [45]

---

[1]The website https://quantumstat.com/dataset/dataset.html has a large list of NLP and QA datasets. The recent survey [67] also mentions many of the datasets.

, Propara [47] , QuaRel [69] , QuAC [13] , MultiRC [33] , Record [85] , Qangaroo [78] , ShARC [62] , SWAG [84] , SQuADv2 [59] in 2018.

To understand the technical challenges and nuances in building natural language understanding (NLU) and QA systems we consider a few examples from the various datasets. In his 1972 paper Winograd [81] presented the following "councilmen" example.

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

In this example to understand what "they" refers to one needs knowledge about the action of "refusing a permit to demonstrate"; when such an action can happen and with respect to whom?

The bAbI domain is a collection of 20 tasks and one of the harder task sub-domain in bAbI is the Path finding sub-domain. An example from that sub-domain is as follows:

> **Context:** The office is east of the hallway. The kitchen is north of the office. The garden is west of the bedroom. The office is west of the garden. The bathroom is north of the garden.
> **Question:** How do you go from the kitchen to the garden?
> **Answer:** South, East.

To answer the above question one needs knowledge about directions and their opposites, knowledge about the effect of actions of going in specific directions, and knowledge about composing actions (i.e., planning) to achieve a goal.

Processes are actions with duration. Reasoning about processes, where they occur and what they change is somewhat more complex. The ProPara dataset [47] consists of natural text about processes. Following is an example.

> **Passage:** Chloroplasts in the leaf of the plant trap light from the sun. The roots absorb water and minerals from the soil. This combination of water and minerals flows from the stem into the leaf. Carbon dioxide (CO_2) enters the leaf. Light, water and minerals, and CO_2 all combine into a mixture. This mixture forms sugar (glucose) which is what the plant eats.
> **Question**: Where is sugar produced? **Answer**: in the leaf

The knowledge needed in many question answering domain can be present in unstructured textual form. But often specific words and phrases in those texts have "deeper meaning" that may not be easy to learn from examples. Such a situation arises in the LifeCycleQA domain [52] where the description of a life cycle is given and questions are asked about them. One LifeCycleQA domain has the description of

the life cycle of frog with five different stages: egg, tadpole, tadpole with legs, froglet and adult. It has description about each of these stages such as:

> **Lifecycle Description:** Froglet - In this stage, the almost mature frog breathes with lungs and still has some of its tail. Adult - The adult frog breathes with lungs and has no tail (it has been absorbed by the body).
> **Question:** What best indicates that a frog has reached the adult stage?
> **Options:** A) When it has lungs B) *When the tail is absorbed by the body*

To answer this question one needs a precise understanding (or definition) of the word "indicates". In this case, (A) indicates that a frog has reached the adult stage (B) does not. That is because (B) is true for both adult and froglet stages while (A) is only true for the adult stage. The precise definition of "indicates" is that a property P indicates a stage S_i with respect to a set of stages $\{S_1, \ldots, S_n\}$ iff P is true in S_i, and P is false in all stages in $\{S_1, \ldots, S_{i-1}, S_{i+1}, \ldots S_n\}$.

There are some datasets where it is explicitly stated that external knowledge is needed to answer questions in those datasets. Example of such datasets are the OpenBookQA dataset [45] and the Machine Commonsense datasets [65; 7; 6]. Each OpenBookQA item has a question and four answer choices. An open book of science facts are also provided. In addition QA systems are expected to use additional common knowledge about the domain. Following is an example item from OpenBookQA.

> **OpenBook Question:** In which of these would the most heat travel?
> **Options:** A) a new pair of jeans. B) *a steel spoon in a cafeteria*. C) a cotton candy at a store. D) a calvin klein cotton hat.
> **Science Fact:** Metal is a thermal conductor
> **Common Knowledge:** Steel is made of metal. Heat travels through a thermal conductor.

As part of the DARPA MCS (machine commonsense) program Allen AI has developed 5 different QA datasets where reasoning with common sense knowledge (which are not given) is required to answer questions correctly. One of those datasets is the Physical QA dataset and following is an example from that dataset.

> **Physical QA:** You need to break a window. Which object would you rather use?
> **Answer Options:** A) *metal stool* B) bottle of water.

To answer the above question one needs knowledge about things that can break a window.

One of the most challenging natural language QA task is solving grid puzzles [50]. They contain a textual description of the puzzle and questions about the solution of the puzzle. Such puzzles are currently found in exams such as LSAT and GMAT, and have also been used to evaluate constraint satisfaction algorithms and systems. Building a system to solve them is a challenge as it requires precise understanding of several clues of the puzzle, leaving very little room for error. In addition there is often need for external knowledge (including commonsense knowledge) that is not explicitly specified in the puzzle description.

> **Puzzle Context:** Waterford Spa had a full appointment calendar booked today. Help Janice figure out the schedule by matching each masseuse to her client, and determine the total price for each.
> **Puzzle Conditions:** Hannah paid more than Teri's client. Freda paid 20 dollars more than Lynda's client. Hannah paid 10 dollars less than Nancy's client. Nancy's client, Hannah and Ginger were all different clients. Hannah was either the person who paid $180 or Lynda's client.

The above examples are a small sample of examples from recent NLQA datasets that illustrate the need for "reasoning" with external knowledge in NLQA. In the rest of the paper we present several aspects of building NLQA systems for such datasets and give a brief survey of the existing methods. We start with a brief description of the knowledge repositories used by such NLQA systems and how they were created. We then present models and architectures of systems that do NLQA with external knowledge.

2 Knowledge Repositories and their creation

The knowledge repositories contain two kinds of knowledge: Unstructured and Structured Knowledge. Unstructured knowledge is knowledge in the form of free text or natural language. Structured knowledge, on the other hand, has a well-defined form, such as facts in the form of tuples or Resource Description Framework (RDF); or well-defined rules as used in Answer Set Program Knowledge Bases. Recently pretrained language models are also being used as "knowledge bases". We discuss them in a later section.

Repositories of Unstructured Knowledge: Any natural language text or book or even the web can be thought of as a source of unstructured knowledge. Two large repositories that are commonly used are the Wikipedia Corpus and the Toronto BookCorpus. Wikipedia corpus contains 4.4M articles about varied fields and is crowd-curated. Toronto BookCorpus consists of 11K books on various topics. Both of these are used in several latest neural language models to learn word representations and NLU.

A few other notable sources of unstructured commonsense knowledge are the Aristo Reasoning Challenge (ARC) Corpus [16], the WikiHow Text Summarization dataset [35], and the short stories from the RoCStories and Story Cloze Task datasets [54]. ARC Corpus contains 14M science sentences which are useful for the corresponding ARC Science QA task. WikiHow dataset contains 230K articles and summaries extracted from the online WikiHow website. This contains articles written by several different human authors, which is different from the news articles present in the Penn Treebank. RocStories and the Story Cloze dataset contains short stories that have a "rich causal and temporal commonsense relations between daily events".

Repositories of Structured Knowledge: There are several structured knowledge repositories, such as Yago [60] , NELL [10] , DBPedia [2] and ConceptNet [40] . Out of these, Yago is human-verified with the confidence of 95% accuracy in its content. NELL on the other hand, continuously learns and updates new facts into its knowledge base. DBPedia is the structured variant of Wikipedia, with facts and relations present in RDF format. ConceptNet contains commonsense data and is an amalgamation of multiple sources, which are DBPedia, Wiktionary, OpenCyc to name a few. OpenCyc is the open-source version of Cyc, which is a long-living AI project to build a comprehensive ontology, knowledge base and commonsense reasoning engine. Wiktionary [80] is a multilingual dictionary describing words using definitions and descriptions with examples. WordNet [24] contains much more lexical information compared to Wiktionary, including words

being grouped into sets of cognitive synonyms expressing a distinct concept. These sets are interlinked by conceptual-semantic and lexical relations.

ATOMIC [64] , VerbPhysics [25] and WebChild [73] are recent collections of commonsense knowledge. WebChild is automatically created by extracting and disambiguating Web contents into triples of nouns, adjectives and relations like *hasShape*, *evokesEmotion*. VerbPhysics contains knowledge about physical phenomena. ATOMIC contains commonsense knowledge, that is curated from human annotators and focuses on inferential knowledge about a given event, such as, intentions of actors, attributes of actors, effects on actors, pre-existing conditions for an event and reactions of actors. It also contains relations, such as, *If-Event-Then-Event*, *If-Event-Then-MentalState* and *If-Event-Then-Persona*.

3 Reasoning with external knowledge: Models and Architectures

The first step in developing a QA system that incorporates reasoning with external knowledge is to have one or more ways to get the external knowledge.

3.1 Extracting the External Knowledge

Neural Language Models: Large neural models, such as BERT [19], that are trained on Masked Language Modeling (MLM) seem to encapsulate a large volume of external knowledge. Their use led to strong performance in multiple datasets and the tasks in GLUE leader board [75]. Recently, methods have been proposed that can extract structured knowledge from such models and are being used for relation-extraction and entity-extraction by modifying MLM as a "fill-in-the-blank" task [22; 9; 8]. These neural models can also be used to generate sentence vector representations, which are used to find sentence similarity and knowledge ranking.

Word Vectors: Even earlier word vectors such as Word2Vec [46] and Glove [55] were used for extracting knowledge. These word vectors imbibe knowledge about word similarities. These word similarities could be used for creating rules that can be used with other handcrafted rules in a symbolic logic reasoning framework. For example, [5] uses word similarities from word vectors and weighted rules in Probabilistic Soft Logic (PSL) to solve the task of Semantic Textual Similarity.

Information/Knowledge Retrieval: However, the task of open-domain NLQA, such as in OpenBookQA mentioned earlier, need external unstructured knowledge in the form of free text. The text is first processed using techniques aimed at removing noise and making retrieval more accurate, such as removal of punctuation and stop-words, lower-casing of words and lemmatization. The processed text is then stored in reverse-indexed in-memory storage such as Lucene [43] and retrieval servers like ElasticSearch [28]. The knowledge retrieval step consists of search keyword identification, initial retrieval from the search engine and then depending on the task, a knowledge re-ranking step. Some complex tasks which require multi-hop reasoning, perform additional such knowledge retrieval steps. In these tasks, the keyword identification step is modified to account for the knowledge sentences retrieved in the previous steps. Multiple such queries make the task computationally expensive.

Few of the structured knowledge sources such as ATOMIC and ConceptNet also contain natural language text examples. These are used as unstructured knowledge and retrieved using the above techniques.

Semantic Knowledge Ranking/Retrieval: The knowledge sentences retrieved through IR are re-ranked further using Semantic Knowledge Ranking/Retrieval (SKR) models, such as in [56; 4; 49; 3]. In SKR, instead of traditional retrieval search engines, neural networks are used to rank knowledge sentences. These neural networks are trained on the task of semantic textual similarity (STS), knowledge relevance classification or natural language inference (NLI).

Existing structured knowledge sources: ConceptNet, WordNet and DBPedia are examples of structured knowledge bases, where knowledge is stored in a well-defined schema. This schema allows for more precise knowledge retrieval systems. On the other hand, the knowledge query too, needs to be well defined and precise. This is the challenge for knowledge retrieval from structured knowledge bases. In these systems, the question is parsed to a well-defined structured query that follows the strict semantics of the target knowledge retrieval system. For example, DBPedia has knowledge stored in RDF triples. A question parser translates a natural language question to a SPARQL query which is executed by the DBPedia Server.

Hand coded knowledge: However, in domains such as the LifeCycleQA where precise definitions of certain terms are needed, such knowledge is hand coded. Similarly, in the puzzle solving domain, ASP rules encoding the fundamental assumption about such puzzles are hand-crafted. One such assumption is that, associations between the elements are unique. The ASP rules for this is:

```
1 {tuple(G,Cat,Elem): element(Cat,Elem) } 1 :- cindex(Cat), eindex(G).
:- tuple(G1,Cat,Elem), tuple(G2,Cat,Elem),G1!=G2.
```

Knowledge learned using Inductive Logic Programming (ILP): Sometimes logical ASP rules can be learned from the dataset. Such a method is used in [52; 48] to learn rules that are used to solve bAbI, ProPara and Math word problem datasets. There, ILP (inductive logic programming) is used to learn effect of actions, and static causal connection between properties of the world. Following is an example of a rule that is learned.

```
initiatedAt(carry(A, O), T ) :- happensAt(take(A, O), T ).
```

The rule says that A is carrying O gets initiated at time point T if the event take(A,O) occurs at T.

3.2 Models and architecture for NLQA with external knowledge

NLQA systems with external knowledge can be grouped based on how knowledge is expressed (structured, free text, implicit in pre-trained neural networks, or a combination) and the type of reasoning module (symbolic, neural or mixed).
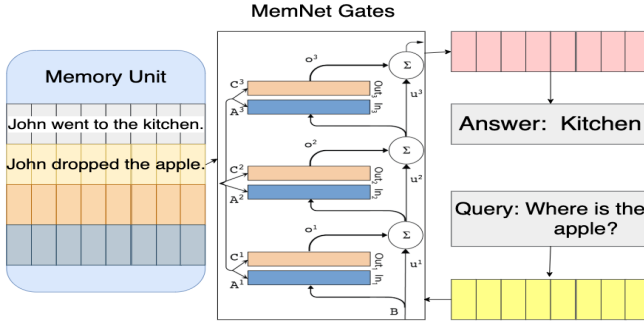
Figure 1: Memory Network with Free text Knowledge. This is an example from the bAbI dataset. The memory unit is first updated from reading the knowledge passage and finally a query is given.

> **Answer Set Programs:** An answer set program (ASP) is a collection of rules of the form, $L_0 \leftarrow L_1, \ldots, L_m, \; not \; L_{m+1}, ..., \; not \; L_n$, where each of the L_i is a literal in the sense of a classical logic. Intuitively, the above rule means that if $L_1, \ldots, L_m$ are true and if $L_{m+1}, \ldots, L_n$ can be safely assumed to be false then L_0 must be true. The semantics of ASP is based on the stable model (answer set) semantics of logic programming. ASP is one of the declarative knowledge representation and reasoning language with a large building block results, efficient interpreters and large variety of applications.

Structured Knowledge and Symbolic Reasoner: While very early NLQA systems were symbolic systems, and most recent NLQA systems are neural, there are a few symbolic systems that do well on the recent datasets. As mentioned earlier the system [5] uses probabilistic soft logic to address semantic textual similarity. The system in [51] learns logical rules from the bAbI dataset using ILP and then does reasoning using ASP and performs well on that dataset.

> **Inductive Logic Programming:** Inductive Logic Programming (ILP) is a subfield of Machine learning that is focused on learning logic programs. Given a set of positive examples E^+, negative examples E^- and some background knowledge B, an ILP algorithm finds an Hypothesis H (answer set program) such that $B \cup H \models E^+$ and $B \cup H \not\models E^-$. The possible hypothesis space is often restricted with a language bias that is specified by a series of mode declarations. Recently, ILP methods have been developed with respect to Answer Set Programming.

The system [48] addresses the Propara dataset in a similar manner and performs well. The system to solve puzzles is also symbolic and so far there has not been any neural system addressing this dataset. However none of these systems are end-to-end symbolic systems as they use neural models for tasks (in the overall pipeline) such as semantic parsing of text.

> **Transformers:** Generative Pre-trained Transformer(GPT) [58] showed that, pretraining on diverse unstructured knowledge texts using transformers [74] and then selectively fine-tune on specific tasks, helps achieving state-of-the-art systems. Their unidirectional way learning representations from diverse texts was improved by bidirectional approach in BERT. It uses the transformer encoder as part of their architecture and takes a fixed length of input and every token is able to attend to both its previous and next tokens through the self attention layers of the transformers. Robustly optimized BERT approach, RoBERTa [42] is another variant which uses a different pre-training approach. RoBERTa is trained on ten times more data than BERT and it outperforms BERT in almost all the NLP tasks and benchmarks.

Neural Implicit Knowledge with Neural Reasoners:

Neural network models based on transformers are able to use the knowledge embedded in their learned parameters in various downstream NLP tasks. These models learn the parameters as a part of *pre-training* phase on a huge collection of diverse free texts. They fine-tune the parameters for selective NLP tasks and in the process refine the pre-trained knowledge even further for a domain and a task. This knowledge in the parameters of the models helps then in multiple NLP tasks. BERT learns the knowledge from mere 16GB, while Roberta uses 160GB of free texts for pre-training. We can say that the knowledge learned during the pre-training phase is generic for any domain and for any NLP task which gets tuned for a particular domain and task during the fine-tuning phase.

Structured Knowledge and Neural Reasoners: Structured knowledge can be in the form of trees (abstract syntax tree, dependency tree, constituency tree), graphs, concepts or rules. Tree-structured knowledge extracted from input text have been used in the Tree-based LSTM [70]. Here, aggregated knowledge from multiple child nodes selected through the gating mechanism in the model propagates to the parent nodes and creates a representation of the whole tree and thus improving the downstream NLP tasks.

> **LSTMs:** Long-Short Term Memory [29] is a type of Recurrent Neural Networks that are used to deal with variable-length sequence inputs. Reasoning and QA with LSTMs are usually modeled as an encoder-classifier architecture. The encoder network is used to learn a vector representation of the question and external knowledge. The classifier network either classifies word-spans to extract answer or answer options to in a MCQA setting. These models are able to encode variable-length sentences and hence can take as input a considerable amount of external knowledge. The limitations of LSTMs are that they are computationally expensive and have limited memory, i.e, they can only remember last T time-steps.

While knowledge in the form of undirected graphs are mainly used by the graph-based reasoning systems like Graph Neural Networks (GNN) [66], the directed graphs are better handled by convolutional neural networks with dense graph propagation [32]. If the knowledge is heterogeneous in nature, having multiple types of objects(graph nodes) and the edges carry different semantic information, heterogeneous graph attention networks [77] is a common choice. Figure 2 shows an example. Knowledge carrying information in hierarchy (entity-level, sentence-level, paragraph-level and document-level) makes use of Hierarchical Graph neural network [23] for better performance.

> **Graph Neural Networks:** GNNs [66] have gained immense importance recently because of it higher expressive power and generalizability across various domains. GNN learns the representation of each nodes of the input graphs by aggregating information from its neighbors through message passing [26]. It learns both semantically and structurally rich representation by iterating over its neighbors in successive hops. In spite of their heavy usage across multiple domains, their shallow structure(in terms of the layers of GNN), inability to handle dynamic graphs and scalability are still open problems [86].

Knowledge can be in the form of structured concepts. One such approach was taken in Knowledge-Powered convolutional neural networks [76] where they use concepts from Probase to classify short text. In another approach, K-BERT [41] inject knowledge graph triples along with other text and together learn representation which helps in NLQA and
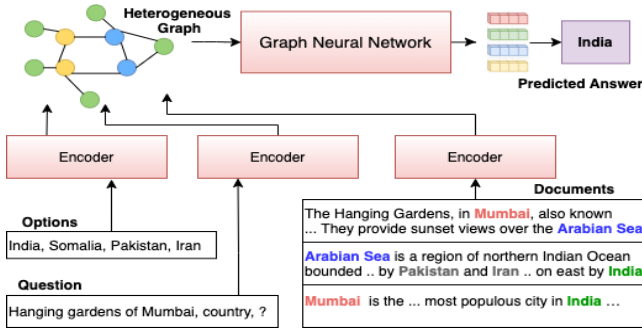
Figure 2: Heterogeneous Graph Neural Network with Structured Knowledge. The example is taken from WikiHop dataset. Based on the documents, question and answer options a heterogeneous graph is created(document nodes in green, option nodes in yellow and entity nodes in blue)

Figure 3: An example from OBQA. Transformer based Neural Models with Free text Knowledge obtained by Information Retrieval on Knowledge Bases.

named entity recognition tasks.

Structured knowledge can be in the form of rules which are compiled into the neural network. One such approach [30] was taken to incorporate first order logic rules into the neural network parameters.

Structured knowledge can also be embedded into a neural network through synthetically created training samples. This can be either template-based or specially hand-crafted to exploit a particular behavior of any model. The knowledge bases such as ConceptNet, WordNet, DBPedia are used to create these hand-crafted training examples. For example, in the NLI task, new knowledge is hand-crafted from these knowledge bases by normalizing the names and switching the roles of the actors. This augmented knowledge helped in differentiating between similar sentences through a symbolic attention mechanism in neural network [53].

Free text Knowledge and Neural Reasoners: Memory networks have been used in prior work on tasks which need long term memory such as the tasks in bAbI datasets [36; 39]. In these models, free text is used as external knowledge and stored in the memory units. The answering module uses this stored memory representation to reason and answer questions.

Memory Networks: Memory Networks augment neural networks with a read and writable *memory* section. These networks are designed to read natural language text and learn to store these input to a specially designed *memory* in which the free text input or a vector representation of the input is written under certain conditions defined by gates. At inference time, a read operation is performed on the memory. These networks are made robust to unseen words during inference with techniques of approximating word vector representations from neighbouring words. These networks outperform RNNs and LSTMs on tasks which require long term memory utilizing both the dedicated *memory* unit and unseen word vectors. Their major limitation lies in their difficulty to be trainable using backpropagation and requiring full supervision during training.

Free text Knowledge in the form of topics, tags and entities have been used along with texts to improve the QA tasks using a knowledge enhanced hybrid neural network [82]. They used knowledge gates to match semantic information of the input text with the knowledge weeding out the unrelated texts.

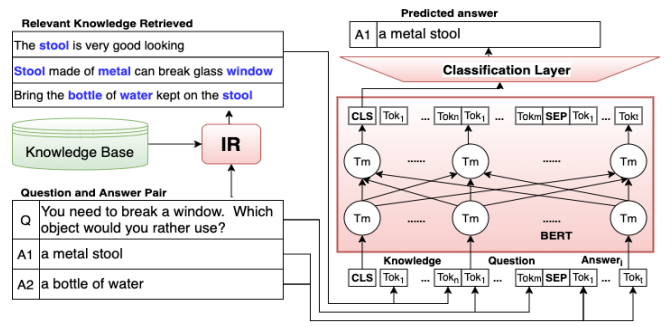In the scenario where the knowledge embedded in parameters of neural models are not enough for an NLP task, external knowledge is often infused in the free text forms along with the model inputs. This has often led to improved reasoning systems. Commonsense reasoning is one such area where external knowledge in the free text form helps in achieving better performance. Multiple challenges has been mentioned earlier which require reasoning abilities.

A common approach for using the free text along with the reasoning systems can be seen from Figure 3. The idea is to pre-select a set of appropriate knowledge texts (using IR) from an unstructured knowledge repository or document and then let the previously learned reasoning models use it [4; 15; 53].

Free text Knowledge and Mixed Reasoners: There have been few approaches in bringing the neural and symbolic approaches together for commonsense reasoning with knowledge represented in text. One such neuro-symbolic approach NeRD [12] was used to achieve state-of-the-art performance in two mathematical datasets DROP and MathQA. NeRD consists of a *reader* which generates representation from questions with passage and a *programmer* which generates the reasoning steps that when executed produces the desired results.

In the LifeCycleQA dataset terms with deeper meaning, such as "indicates", are defined explicitly using Answer Set Programming (ASP). The overall reasoning system in [52] is an ASP program that makes use of such ASP defined terms and performs external calls to a neural NLI system. In [57] the Winograd dataset is addressed by using "extracted sentences from web" and BERT and combined reasoning is done using PSL.

Combination of Knowledge and Mixed Reasoners: The ARISTO solver by AllenAI [16] consists of eight reasoners that utilize a combination of external knowledge sources. The external knowledge sources include Open-IE [21], the ARC corpus, TupleKB, Tablestore and TupleInfKB [15]. The solvers range from retrieval-based and statistical models, symbolic solvers using ILP, to large neural language models such as BERT. They combine the models using two-step ensembling using logistic regression over base reasoner scores.

4 Discussion: How much Reasoning the models are doing?

While it is easy to see the various kinds of reasoning done by symbolic models, it is a challenge to figure out how much reasoning is being done in neural models. The success with respect to datasets give some indirect evidence, but more thorough studies are needed.

Commonsense Reasoning: The commonsense datasets, (CommonsenseQA, CosmosQA, SocialIQA, etc.), require commonsense knowledge to be solved. So the systems solving them should perform deductive reasoning with external commonsense knowledge. BERT-based neural models are able to perform reasonably well on such tasks, with accuracy ranging from 75%-82% which falls quite short of human performance of 90-94% [49]. They leverage both external commonsense knowledge and knowledge learned through pre-training tasks. Inability of the pure neural models to represent such a huge variety of commonsense knowledge as rules led to pure symbolic or neuro-symbolic approaches on such datasets.

Multi-Hop Reasoning: NLQA datasets such as, Natural Questions, HotpotQA, OpenBookQA, etc, need systems to combine information spread across multiple sentences, i.e, Multi-Hop Reasoning. Current state-of-the-art neural models are able to achieve performance around 65-75% on these tasks [4]. Transformer models have shown a considerable ability to perform multi-hop reasoning, particularly for the case where the entire knowledge passage can be given as an input, without truncation. Systems have been designed to select and filter sentences to reduce confusion in neural models

Abductive Reasoning: AbductiveNLI, a dataset from AllenAI, defines a task to abduce and infer which of the possible scenarios best explains the consequent events. Neural models have shown a considerable performance in the simple two-choice MCQ task [49], but have a low performance in the abduced hypothesis generation task. This shows the neural models are able to perform abduction to a limited extent.

Quantitative and Qualitative Reasoning: Datasets such as DROP, AQuA-RAT, Quarel, Quartz require systems to perform both qualitative and quantitative reasoning. The quantitative reasoning datasets like DROP, AQuA-RAT, led to emergence of approaches which can learn symbolic mathematics using deep learning and neuro-symbolic models [38; 12]. Neural as well as symbolic approaches have shown decent performance(70-80%) on qualitative reasoning compared to human performance of (94-95%) [69; 68]. All the datasets have shown a considerable diversity in systems with all three types (neural, symbolic and neuro-symbolic) of models performing well.

Non-Monotonic Reasoning: In [17] an attempt is made to understand the limits of reasoning of Transformers. They note that the transformers can perform simple logical reasoning and can understand rules in natural language which corresponding to classical logic and can perform limited non-monotonic reasoning. Though there are few datasets which require such capabilities.

5 Conclusion and Future Directions

In this paper we have surveyed [2] recent research on NLQA when external knowledge – beyond what is given in the test part – is needed in correctly answering the questions. We gave several motivating examples, mentioned several datasets, discussed available knowledge repositories and methods used in selecting needed knowledge from larger repositories, and analyzed and grouped several models and architectures of NLQA systems based on how the knowledge is expressed in them and the type of reasoning module used in them. Although there have been some related recent surveys, such as [67], none of them focus on how exactly knowledge and reasoning is done in the NLQA systems dealing with datasets that require external knowledge. Our survey touched upon knowledge types that include structured knowledge, textual knowledge, knowledge embedded in a neural network, knowledge provided via specially constructed examples and combinations of them. We explored how symbolic, neural and mixed models process and reason with such knowledge. Based on our observations for various models following are some questions and future directions.

Following up on the LifeCycleQA dataset where some concepts, such as "indicates", were manually defined, several questions – with partial answers – may come to mind. (i) How big is the list of such concepts? If a list of them is made and they are defined then these definitions can be directly used or compiled into neural models. Can Cyc and the book [27] be starting points in this direction? (ii) Can these definitions be learned from data? How? Unless one only focuses on specific datasets, the challenge in learning these definitions would be that for each of them specialized examples would have to be created. (iii) When is it easier to just write the definitions in a logical language? When is it easier to learn from data sets? Some concepts are easy to define logically but may require lot of examples to teach a system. There are concepts whose definitions took decades for researchers to formalize. An example is the solutions of frame problem. So it is easier to use that formalization rather than learning from scratch. On the other hand the notion of "cause" is still being fine tuned. Another direction of future work centers around the question of how well neural network can do reasoning; What kind of reasoning they can do well and what are the challenges ?

We hope this paper encourages further interactions between researchers in the traditional area of knowledge representation and reasoning who mostly focus on structured knowledge, and researchers in NLP/NLU/QA who are interested in NLQA where reasoning with knowledge is important. That would be a next step towards better understanding of natural language and development of cognitive skills. It will help us move from knowledge and comprehension to application, analysis and synthesis, as described in Bloom's taxonomy.

---

[2]A version of this paper with a larger bibliography is linked here. If accepted, we plan to buy extra pages to accommodate them.

References

[1] A Abujabal, R S Roy, M Yahya, and G Weikum. Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters. *arXiv:1809.09528*, 2018.

[2] S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, and Z Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, ISWC'07/ASWC'07, pages 722–735. Springer-Verlag, 2007.

[3] Pratyay Banerjee. ASU at TextGraphs 2019 shared task: Explanation ReGeneration using language models and iterative re-ranking. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for NLP (TextGraphs-13)*, pages 78–84, Hong Kong, November 2019. ACL.

[4] Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. Careful selection of knowledge to solve open book question answering. In *ACL*, pages 6120–6129, Florence, Italy, July 2019. ACL.

[5] Islam Beltagy, Katrin Erk, and Raymond Mooney. Probabilistic soft logic for semantic textual similarity. In *ACL (Volume 1: Long Papers)*, pages 1210–1219, Baltimore, Maryland, June 2014. ACL.

[6] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019.

[7] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.

[8] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. *CoRR*, abs/1906.05317, 2019.

[9] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. Inducing relational knowledge from bert, 2019.

[10] A Carlson, J Betteridge, B Kisiel, B Settles, E R Hruschka, and T M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[11] M Chen, M D'Arcy, A Liu, J Fernandez, and D Downey. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, 2019.

[12] Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song, and Quoc V Le. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In *ICLR*, 2019.

[13] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

[14] C Clark, K Lee, MW Chang, T Kwiatkowski, M Collins, and K Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv:1905.10044*, 2019.

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

[16] Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. From 'f' to 'a' on the n.y. regents science exams: An overview of the aristo project. *ArXiv*, abs/1909.01958, 2019.

[17] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020.

[18] P Dasigi, N F Liu, A Marasovic, N A Smith, and M Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *arXiv:1908.05803*, 2019.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*, 2019.

[21] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[22] A. H. Miller P. Lewis A. Bakhtin Y. Wu F. Petroni, T. Rocktäschel and S. Riedel. Language models as knowledge bases? In *EMNLP*, 2019.

[23] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. Hierarchical graph network for multi-hop question answering. *arXiv preprint arXiv:1911.03631*, 2019.

[24] C Fellbaum. Wordnet. *The encyclopedia of applied linguistics*, 2012.

[25] M. Forbes and Y. Choi. Verb physics: Relative physical knowledge of actions and objects. *arXiv:1706.03799*, 2017.

[26] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[27] A. S. Gordon and J. R. Hobbs. *A formal theory of commonsense psychology: How people think people think.* Cambridge University Press, 2017.

[28] Clinton Gormley and Zachary Tong. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.", 2015.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[30] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *ACL (Volume 1: Long Papers)*, pages 2410–2420, Berlin, Germany, August 2016. ACL.

[31] L Huang, R L Bras, C Bhagavatula, and Y Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv:1909.00277*, 2019.

[32] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496, 2019.

[33] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *NAACL*, 2018.

[34] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. *arXiv preprint arXiv:1910.11473*, 2019.

[35] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

[36] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, pages 1378–1387, 2016.

[37] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[38] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. In *ICLR*, 2020.

[39] Fei Liu and Julien Perez. Gated end-to-end memory networks. In *ACL: Volume 1, Long Papers*, pages 1–10, 2017.

[40] H Liu and P Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[41] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. *arXiv preprint arXiv:1909.07606*, 2019.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[43] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0.* Manning Publications Co., USA, 2010.

[44] J. McCarthy. An example for NLU and the AI problems it raises. *Formalizing Common Sense: Papers by J. McCarthy*, 355, 1990.

[45] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[46] T Mikolov, I Sutskever, K Chen, G Corrado, and J Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[47] B. Mishra and et al. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. *arXiv:1805.06975*, 2018.

[48] Arindam Mitra. *Knowledge Representation, Reasoning and Learning for Non-Extractive Reading Comprehension.* PhD thesis, Arizona State University, 2019.

[49] Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *arXiv preprint arXiv:1909.08855*, 2019.

[50] Arindam Mitra and Chitta Baral. Learning to automatically solve logic grid puzzles. In *EMNLP*, pages 1023–1033, 2015.

[51] Arindam Mitra and Chitta Baral. Addressing a QA challenge by combining statistical methods with inductive rule learning and reasoning. In *AAAI*, 2016.

[52] Arindam Mitra, Peter Clark, Oyvind Tafjord, and Chitta Baral. Declarative question answering over knowledge bases containing natural language text with answer set programming. In *AAAI*, volume 33, pages 3003–3010, 2019.

[53] Arindam Mitra, Ishan Shrivastava, and Chitta Baral. Understanding roles and entities: Datasets and models for natural language inference. *arXiv preprint arXiv:1904.09720*, 2019.

[54] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*, 2016.

[55] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[56] George Sebastian Pirtoaca, Traian Rebedea, and Stefan Ruseti. Answering questions by learning to rank-learning to rank by answering questions. In *EMNLP-IJCNLP*, pages 2531–2540, 2019.

[57] Ashok Prakash, Arpit Sharma, Arindam Mitra, and Chitta Baral. Combining knowledge hunting and neural language models to solve the winograd schema challenge. In *ACL*, pages 6110–6119, 2019.

[58] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[59] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[60] T Rebele, F Suchanek, J Hoffart, J Biega, E Kuzey, and G Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *ISWC*, pages 177–185. Springer, 2016.

[61] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[62] M Saeidi, M Bartolo, P S. H. Lewis, S Singh, T Rocktäschel, M Sheldon, G Bouchard, and S Riedel. Interpretation of natural language rules in conversational machine reading. *CoRR*, abs/1809.01494, 2018.

[63] K Sakaguchi, R L Bras, C Bhagavatula, and Y Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv:1907.10641*, 2019.

[64] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: an atlas of machine commonsense for if-then reasoning. In *AAAI*, volume 33, pages 3027–3035, 2019.

[65] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[66] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

[67] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

[68] O Tafjord, M Gardner, K Lin, and P Clark. Quartz: An open-domain dataset of qualitative relationship questions. *arXiv:1909.03553*, 2019.

[69] Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. Quarel: A dataset and models for answering questions about qualitative relationships. *arXiv preprint arXiv:1811.08048*, 2018.

[70] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[71] A Talmor and J Berant. The web as a knowledge-base for answering complex questions. *arXiv:1803.06643*, 2018.

[72] A Talmor, J Herzig, N Lourie, and J Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv:1811.00937*, 2018.

[73] N. Tandon, G. De Melo, F. Suchanek, and G. Weikum. Webchild: Harvesting and organizing commonsense knowledge from the web. In *WSDM*, pages 523–532, 2014.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[75] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. ACL.

[76] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, pages 2915–2921, 2017.

[77] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The World Wide Web Conference*, pages 2022–2032, 2019.

[78] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

[79] J. Weston and et al. Towards AI-complete QA. *arXiv:1502.05698*, 2015.

[80] Wikipedia contributors. Wikipedia, the free encyclopedia, 2004. [Online; accessed 22-July-2004].

[81] T. Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[82] Yu Wu, Wei Wu, Can Xu, and Zhoujun Li. Knowledge enhanced hybrid neural network for text matching. In *AAAI*, 2018.

[83] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380, 2018.

[84] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

[85] S Zhang, X Liu, J Liu, J Gao, K Duh, and B Van Durme. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv:1810.12885*, 2018.

[86] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.