Question Answering and Machine Reading Comprehension

Jiao Sun
Computer Science Department
University of Southern California
jiaosun@usc.edu

1 Introduction

Recent years have witnessed an increasing demand for conversational Question Answering (QA) agents that allow users to query a large-scale Knowledge Base (KB) or a document collection in natural language. The former is known as KB-QA agents and the latter text-QA agents. In the next section, we will follow the line of a review of KB -> symbolic approaches to KB-QA based on semantic parsing -> multi-step reasoning on KB -> review some KB-QA datasets developed recently.

Then we will talk about text-QA. The heart of it is a neural Machine Reading Comprehension (MRC) model that generates an answer to an input question. The guideline will be: reviewing some MRC datasets -> MRC models -> analysis of these methods -> new datasets addressing the existing problems.

2 Knowledge Base and KB-QA

Review KB. Organizing the world's facts and storing them in a structured database, large scale Knowledge Bases (KB) like DBPedia, Freebase and Yago have become important resources for supporting open-domain QA.

A typical KB consists of a collection of subject-predicate-object triples (s, r, t) where s, t are entities and r is a predicate or relation. A KB in this form is often called a knowledge graph due to its graphical representation.

Semantic Parsing for KB-QA. Most state-of-the-art symbolic approaches to KB-QA are based on semantic parsing, where a question is mapped to its formal meaning representation (e.g., logical form) and then translated to a KB query. The answers to the question can then be obtained by finding a set of paths in the KB that match the query and retrieving the end nodes of these paths.

We take the example used in Yih et al. [2015] to illustrate the QA process. Figure 1(right) shows the logical form in $\lambda-$calculus and its equivalent graph representation. The query graph is grounded in Freebase.

A symbolic KB-QA system to a very large KB is challenging for two reasons:

- **Paraphrasing in natural language:** We need to measure how likely the predicate used in the question matches that in the Freebase, such as "who first voiced Meg on Family Guy" vs. "cast-actor". Embedding-based methods are proposed to solve this problem. The bilinear model in Yang et al. [2014], Nguyen [2017] is one of the basic KB embedding models. It learns a vector x_e for each entity e and a matrix W_r for each relation. The model scores how likely a triple (s, r, t) holds using $score(s, r, t; \theta) = x_s^T W_r x_t$. The loss is defined as $L(\theta) = \sum(\gamma + score(x^-; \theta - score(x^+; \theta)))$. Interested users are referred to Nguyen [2017] for a detailed survey of embedding models for KBC.

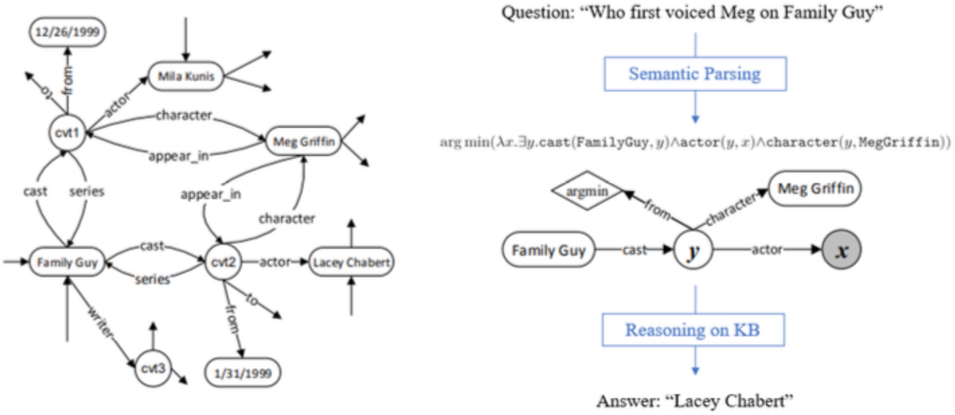


Figure 1: An example of semantic parsing for KB-QA. (Left) A subgraph of Freebase related to the TV show Family Guy. (Right) A question, its logical form in λ-calculus and query graph, and the answer.

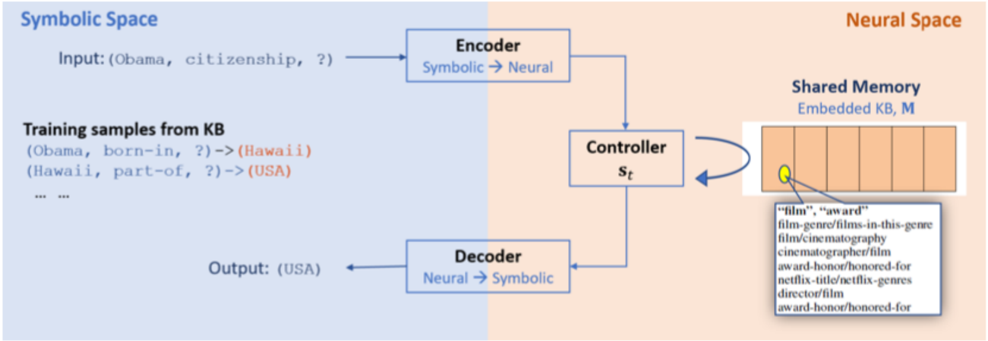
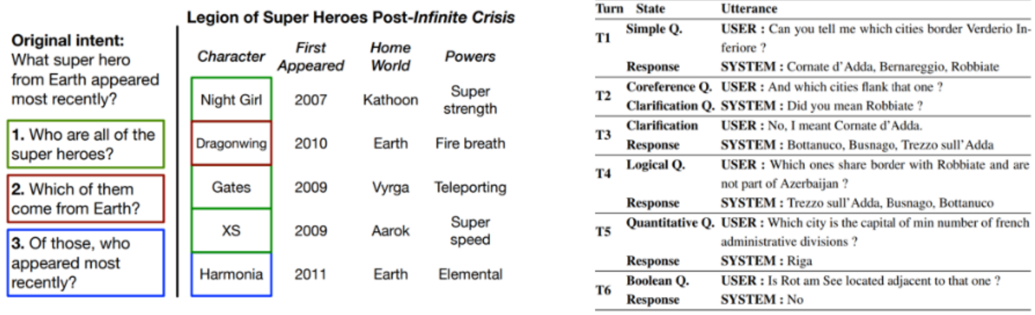


Figure 2: An overview of the neural methods for KBR (Shen et al., 2017a; Yang et al., 2017a). The KB is embedded in neural space as matrix M that is learned to store compactly the connections between related triples (e.g., the relations that are semantically similar are stored as a cluster). The controller is designed to adaptively produce lookup sequences in M and decide when to stop.

- **Search complexity.** Searching all possible multi-step relation paths that match complex queries is expensive. We will introduce more about traversing the query-matched paths in a KB to find the answers.

Multi-Step Reasoning on KB. To reason over a KB, for each relation r, we are interested in learning a first-order logical rules. Given the question "What is citizenship of Obama?" its translated KB query is (Obama, citizenship, ?), a multi-step reasoning is needed to deduce the answer from the paths that contain relevant triples, such as (Obama, born-in, Hawaii) and (Hawaii, part-of, USA), using the learned relational paths such as (born-in, part-of). There are three categories of multi-step KBR methods. They differ in whether reasoning is performed in a discrete symbolic space or a continuous space.

- **Symbolic methods.** The path ranking algorithm [Lao et al., 2011] is one of the primary symbolic approaches to learning relational paths in large KBs. During KBR, given a query q=(s,r,?), PRA selects the set of relational paths for r, denoted by $\pi_1, \pi_2, ...$, then traverse the KB according to the KB according to the query and path, and scores each candidate answer t using a linear model $score(q,t) = \sum \lambda_\pi P(t|s,\pi)$, where λ_{pi} are the learned weights.


Original intent:
What super hero from Earth appeared most recently?
1. Who are all of the super heroes?
2. Which of them come from Earth?
3. Of those, who appeared most recently?
Legion of Super Heroes Post-*Infinite Crisis*

Character	First Appeared	Home World	Powers
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

Turn	State	Utterance
T1	**Simple Q.**	**USER :** Can you tell me which cities border Verderio Inferiore ?
	Response	**SYSTEM :** Cornate d'Adda, Bernareggio, Robbiate
T2	**Coreference Q.**	**USER :** And which cities flank that one ?
	Clarification Q.	**SYSTEM :** Did you mean Robbiate ?
T3	**Clarification**	**USER :** No, I meant Cornate d'Adda.
	Response	**SYSTEM :** Bottanuco, Busnago, Trezzo sull'Adda
T4	**Logical Q.**	**USER :** Which ones share border with Robbiate and are not part of Azerbaijan ?
	Response	**SYSTEM :** Trezzo sull'Adda, Busnago, Bottanuco
T5	**Quantitative Q.**	**USER :** Which city is the capital of min number of french administrative divisions ?
	Response	**SYSTEM :** Riga
T6	**Boolean Q.**	**USER :** Is Rot am See located adjacent to that one ?
	Response	**SYSTEM :** No

Figure 3: The examples from two conversational KB-QA datasets. (Left) An example question sequence created from a compositional question intent in the SQA dataset. (Right) An example dialogue from the CSQA dataset.

Because PRA operates in a fully discrete space, it not only hurts generalization but makes reasoning expensive.

- **Neural Methods.** Implicit ResoNet(IRN) and Neural Logic Programming are proposed to perform multi-step KBR in a neural space. The basic structure is shown in
- **Reinforcement Learning based Methods.** DeepPath, MINERVA and M-Walk are among the state-of-the-art methods that use RL for multi-step reasoning over a KB. They use a policy-based agent with continuous states to traverse the knowledge graph to identify the answer node for an input query.

Datasets. Sequential Question Answering [Iyyer et al., 2017] is collected via crowdsourcing by leveraging WikiTableQuestions(WTQ), which contains highly compositional questions associated with HTML tables from Wikipedia. The workers are asked to compose a sequence of simpler but inter-related questions that lead to the final intent. The answers to the simple questions are subsets of the cells in the table as shown in the left side of Figure 3.

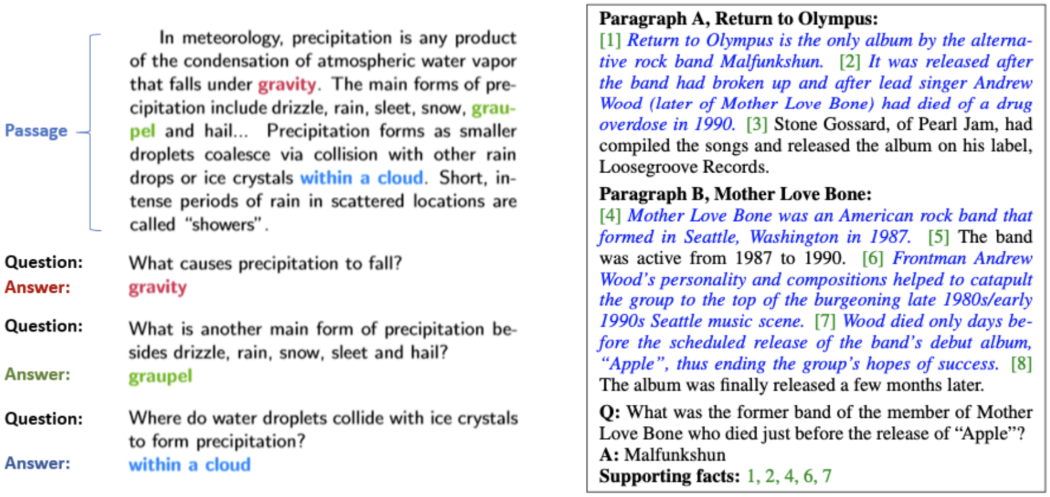
Saha et al. [2018] presented a dataset consisting of 200K QA dialogues for the task of complex sequential question answering as shown in the right side of Figure 3.

3 Machine Reading for Text-QA

Machine Reading Comprehension (MRC) is a challenging task: the goal is to have machines read a (set of) text passage(s) and then answer any question about the passage(s).

The recent big progress on MRC is largely due to the availability of a multitude of large-scale datasets that the research community has created over various text sources such as Wikipedia (WikiReading (Hewlett et al., 2016), SQuAD (Rajpurkar et al., 2016), WikiHop (Welbl et al., 2017), DRCD (Shao et al., 2018)), news and other articles (CNN/Daily Mail (Hermann et al., 2015), NewsQA (Trischler et al., 2016), RACE (Lai et al., 2017), ReCoRD (Zhang et al., 2018)), fictional stories (MCTest (Richardson et al., 2013), CBT (Hill et al., 2015), science questions (ARC (Clark et al., 2018)), and general Web documents (MS MARCO (Nguyen et al., 2016), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), DuReader (He et al., 2017b), QAngaroo [Welbl et al., 2018], HotpotQA [Yang et al., 2018] and etc. HotpotQA is getting more attention because the pervious datasets have some limits:

- In SQuAD questions are designed to be answered given a single paragraph as the context, and most of the questions can be answered by matching the question with a single sentence in that paragraph.
- TrivaQA and SearchQA collect multiple documents to from the context. However, most of the questions can be answered by matching the question with a few nearby sentences in one single paragraph, which is limited as it does not require more complex reasoning.



Passage

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

Question: What causes precipitation to fall?
Answer: gravity

Question: What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
Answer: graupel

Question: Where do water droplets collide with ice crystals to form precipitation?
Answer: within a cloud

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 4: The examples from two MRC datasets. (Left) Question-answer pairs for a sample passage in the SQuAD dataset. (Right)An example of the multi-hop questions in HOTPOTQA. They also highlight the supporting facts in blue italics, which are also part of the dataset.

- Target multi-hop reasoning: QAngaroo and COMPLEXWEBQUESTIONS are constructed using existing knowledge bases, and thus are constrained by the schema of KBs they use.
- All of above datasets only provide distant supervision, i.e., the systems only know the what the answer is, but do not know what the leading facts.

4 MRC Models and Analysis

Given a question $Q = (q_1, ..., q_I)$ and a passage $P = (p_1, ..., p_J)$, we need to locate an answer span $A = (a_start, a_end)$ in P. Most Neural MRC models encode questions and passages through three layers: a lexicon embedding layer, a contextual embedding layer, and an attention layer, as reviewed below.

Lexicon Embedding Layer. This extracts information from Q and P at the word level and normalizes for lexical variants. It typically maps each word to a vector space using a pre-trained word embedding model, such as word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), such that semantically similar words are mapped to the vectors that are close to each other in the neural space.

Contextual Embedding Layer. This utilizes contextual cues from surrounding words to refine the embedding of the words. As a result, the same word might map to different vectors in a neural space depending on its context, such as "bank of a river" vs. " bank of America". This is typically achieved by using a Bi-directional Long Short-Term Memory (BiLSTM) network.

ELMo [Peters et al., 2018] is one of the state of the art contextual embedding models. It is based on deep BiLSTM. Instead of using only the output layer representations of BiLSTM, ELMo combines the intermediate layer representations in the BiLSTM, where the combination weights are optimized on task-specific training data.

BERT [Devlin et al., 2019] differs from ELMo and BiLSTM in that it is designed to pre-train deep bidirection representations by jointly conditioning on both left and right context in all layers. The pre-trained BERT representations can be fine-tuned with just one additional output layer to create state of the art models for a wide range of NLP tasks, including MRC.

Attention Layer. Attention has been a highly successful neural network component. It is naturally interpretable because an attention weight has a clear meaning: how much a particular word will be weighted when computing the next representation for the current word.

Claim	Google is not a harmful monopoly
Reason	People can choose not to use Google
Warrant	Other search engines don't redirect to Google
Alternative	All other search engines redirect to Google

Reason (and since) **Warrant** → **Claim**
Reason (but since) **Alternative** → ¬ **Claim**

	Test		
	Mean	**Median**	**Max**
BERT	**0.671** ± 0.09	**0.712**	**0.770**
BERT (W)	0.656 ± 0.05	0.675	0.712
BERT (R, W)	0.600 ± 0.10	0.574	0.750
BERT (C, W)	0.532 ± 0.09	0.503	0.732
BoV	0.564 ± 0.02	0.569	0.595
BoV (W)	0.567 ± 0.02	0.572	0.606
BoV (R, W)	0.554 ± 0.02	0.557	0.579
BoV (C, W)	0.545 ± 0.02	0.544	0.589
BiLSTM	0.552 ± 0.02	0.552	0.592
BiLSTM (W)	0.550 ± 0.02	0.547	0.577
BiLSTM (R, W)	0.547 ± 0.02	0.551	0.577
BiLSTM (C, W)	0.552 ± 0.02	0.550	0.601

Figure 5: Left: an example of a data point from the Argument Reasoning Comprehension Task; Right: Results of probing experiments with BERT, and the BoV and BiLSTM baselines. These results indicate that BERT's peak 77% performance can be entirely accounted for by exploiting spurious cues.By just considering warrants (W) we can get to 71%.

Heuristic	Premise	Hypothesis	Label
Lexical overlap heuristic	The banker near the judge saw the actor.	The banker saw the actor.	E
	The lawyer was advised by the actor.	The actor advised the lawyer.	E
	The doctors visited the lawyer.	The lawyer visited the doctors.	N
	The judge by the actor stopped the banker.	The banker stopped the actor.	N
Subsequence heuristic	The artist and the student called the judge.	The student called the judge.	E
	Angry tourists helped the lawyer.	Tourists helped the lawyer.	E
	The judges heard the actors resigned.	The judges heard the actors.	N
	The senator near the lawyer danced.	The lawyer danced.	N
Constituent heuristic	Before the actor slept, the senator ran.	The actor slept.	E
	The lawyer knew that the judges shouted.	The judges shouted.	E
	If the actor slept, the judge saw the artist.	The actor slept.	N
	The lawyers resigned, or the artist slept.	The artist slept.	N

Figure 6: Examples of sentences used to test the three heuristics. A model relying on the heuristics would label all examples as entailment.

5 BERT is still not smart

Introduce GLUE. Wang et al. [2018] introduced a battery of nine reading-comprehension tasks for computers called GLUE (General Language Understanding Evaluation). The test was designed as a fairly representative sample of what the research community thought were interesting challenges, but pretty straightforward for humans. For example, one task asks whether a sentence is true based on information offered in a preceding sentence. If you can tell that "President Trump landed in Iraq for the start of a seven-day visit" implies that "President Trump is on an overseas visit," you've just passed. Before BERT, the highest score was 69, while BERT scores 80 six months later.

BERT is just learning spurious statistical cues? Niven and Kao [2019] used BERT to achieve an impressive result on a relatively obscure natural language understanding benchmark called the argument reasoning comprehension task. But instead of concluding that BERT could apparently imbue neural networks with near-Aristotelian reasoning skills, they suspected a simpler explanation: that BERT was picking up on superficial patterns in the way the warrants were phrased. Indeed, after re-analyzing their training data, the authors found ample evidence of these so-called spurious cues. For example, simply choosing a warrant with the word "not" in it led to correct answers 61% of the time. After these patterns were scrubbed from the data, BERT's score dropped from 77 to 53 — equivalent to random guessing.

Right for the wrong reasons? McCoy et al. [2019] published evidence that BERT's high performance on certain GLUE tasks might also be attributed to spurious cues in the training data for those tasks. By changing the heuristics [HANS dataset], the performance of BERT and other methods decrease dramatically. BERT gets near-perfect accuracy when the true label is entailment, and near-zero accuracy when the true label is non-entailment.

Deep models such as BERT do not demonstrate robust commonsense reasoning ability by themselves. Instead, they operate more like rapid surface learners By themselves.

Encourage robust understanding. One way to encourage progress toward robust understanding is to focus not just on building a better BERT, but also on designing better benchmarks and training data that lower the possibility of Clever Hans–style cheating. Zellers et al. [2019] use Adversarial Filtering to produce a challenging dataset called HelloSwag. On each iteration, a new classifier is trained on a dummy training set to replace easily-classified negative endings on the dummy test set with adversarial endings.

Wang et al. [2019] recently introduced a test called SuperGLUE specifically designed to be hard for BERT-based systems. So far, no neural network can beat human performance on it.

There is also research [Clark et al., 2019] understanding what BERT's attention looks at. The authors find that early heads attend to [CLS], middle heads attend to [SEP], and deep heads attend to periods and commas. Often more than half of a head's total attention is to these tokens. Attention heads in the same layer tend to behave similarly.

Conclusion. As Sam Bowman said, who came up with GLUE in the first place, "we are definitely in an era where the goal is to keep coming up with harder problems that represent language understanding, and keep figuring out how to solve those problems".

References

K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

M. Iyyer, W.-t. Yih, and M.-W. Chang. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1167. URL `https://www.aclweb.org/anthology/P17-1167`.

N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 529–539. Association for Computational Linguistics, 2011.

T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL `https://www.aclweb.org/anthology/P19-1334`.

D. Q. Nguyen. An overview of embedding models of entities and relationships for knowledge base completion. *arXiv preprint arXiv:1703.08098*, 2017.

T. Niven and H.-Y. Kao. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

A. Saha, V. Pahuja, M. M. Khapra, K. Sankaranarayanan, and S. Chandar. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph, 2018.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *ArXiv*, abs/1905.00537, 2019.

J. Welbl, P. Stenetorp, and S. Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. doi: 10.1162/tacl_a_00021. URL `https://www.aclweb.org/anthology/Q18-1021`.

B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

S. W.-t. Yih, M.-W. Chang, X. He, and J. Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. 2015.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/p19-1472. URL `http://dx.doi.org/10.18653/v1/p19-1472`.