Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks

Rajarshi Das[*♠] **Manzil Zaheer**[*♡] **Siva Reddy**[♣] and **Andrew McCallum**[♠]
[♠]College of Information and Computer Sciences, University of Massachusetts Amherst
[♡]School of Computer Science, Carnegie Mellon University
[♣]School of Informatics, University of Edinburgh
{rajarshi, mccallum}@cs.umass.edu, manzilz@cs.cmu.edu
siva.reddy@ed.ac.uk

Abstract

Existing question answering methods infer answers either from a knowledge base or from raw text. While knowledge base (KB) methods are good at answering compositional questions, their performance is often affected by the incompleteness of the KB. Au contraire, web text contains millions of facts that are absent in the KB, however in an unstructured form. *Universal schema* can support reasoning on the union of both structured KBs and unstructured text by aligning them in a common embedded space. In this paper we extend universal schema to natural language question answering, employing *memory networks* to attend to the large body of facts in the combination of text and KB. Our models can be trained in an end-to-end fashion on question-answer pairs. Evaluation results on SPADES fill-in-the-blank question answering dataset show that exploiting universal schema for question answering is better than using either a KB or text alone. This model also outperforms the current state-of-the-art by 8.5 F_1 points.

1 Introduction

Question Answering (QA) has been a long-standing goal of natural language processing. Two main paradigms evolved in solving this problem: 1) answering questions on a knowledge base; and 2) answering questions using text.

Knowledge bases (KB) contains facts expressed in a fixed schema, facilitating compositional reasoning. These attracted research ever since the early days of computer science, e.g., BASEBALL (Green Jr et al., 1961). This problem has matured into learning semantic parsers from parallel question and logical form pairs (Zelle and Mooney, 1996; Zettlemoyer and Collins, 2005), to recent scaling of methods to work on very large KBs like Freebase using question and answer pairs (Berant et al., 2013). However, a major drawback of this paradigm is that KBs are highly incomplete (Dong et al., 2014). It is also an open question whether KB relational structure is expressive enough to represent world knowledge (Stanovsky et al., 2014; Gardner and Krishnamurthy, 2017)

The paradigm of exploiting text for questions started in the early 1990s (Kupiec, 1993). With the advent of web, access to text resources became abundant and cheap. Initiatives like TREC QA competitions helped popularizing this paradigm (Voorhees et al., 1999). With the recent advances in deep learning and availability of large public datasets, there has been an explosion of research in a very short time (Rajpurkar et al., 2016; Trischler et al., 2016; Nguyen et al., 2016; Wang and Jiang, 2016; Lee et al., 2016; Xiong et al., 2016; Seo et al., 2016; Choi et al., 2016). Still, text representation is unstructured and does not allow the compositional reasoning which structured KB supports.

An important but under-explored QA paradigm is where KB and text are exploited together (Ferrucci et al., 2010). Such combination is attractive because text contains millions of facts not present in KB, and a KB's generative capacity represents infinite number of facts that are never seen in text. However QA inference on this combination is challenging due to the structural non-uniformity of KB and text. *Distant supervision* methods (Bunescu and Mooney, 2007; Mintz et al., 2009; Riedel et al., 2010; Yao et al., 2010; Zeng et al., 2015) address this problem partially by means of aligning text patterns with KB. But the rich and ambiguous nature of language allows a fact to be expressed in many different forms which these models fail to capture. *Universal schema* (Riedel et al., 2013) avoids the alignment problem by jointly embedding KB facts

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 358–365
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2057

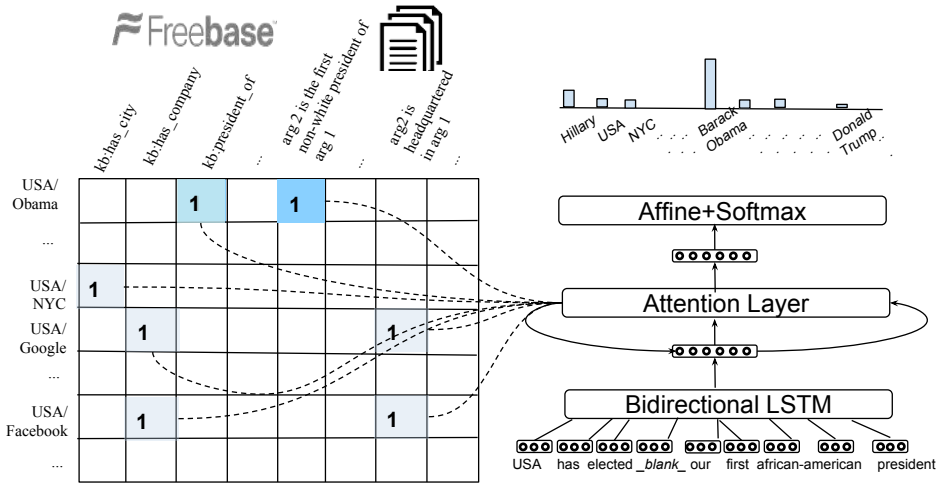

Figure 1: Memory network attending the facts in the universal schema (matrix on the left). The color gradients denote the attention weight on each fact.

and text into a uniform structured representation, allowing interleaved propagation of information. Figure 1 shows a universal schema matrix which has pairs of entities as rows, and Freebase and textual relations in columns. Although universal schema has been extensively used for relation extraction, this paper shows its applicability to QA. Consider the question *USA has elected _blank_, our first african-american president* with its answer *Barack Obama*. While Freebase has a predicate for representing presidents of USA, it does not have one for 'african-american' presidents. Whereas in text, we find many sentences describing the presidency of Barack Obama and his ethnicity at the same time. Exploiting both KB and text makes it relatively easy to answer this question than relying on only one of these sources.

Memory networks (MemNN; Weston et al. 2015) are a class of neural models which have an external memory component for encoding short and long term context. In this work, we define the memory components as observed cells of the universal schema matrix, and train an end-to-end QA model on question-answer pairs.

The contributions of the paper are as follows (a) We show that universal schema representation is a better knowledge source for QA than either KB or text alone, (b) On the SPADES dataset (Bisk et al., 2016), containing real world fill-in-the-blank questions, we outperform state-of-the-art semantic parsing baseline, with 8.5 F_1 points. (c) Our analysis shows how individual data sources help fill the weakness of the other, thereby improving overall performance.

2 Background

Problem Definition Given a question q with words $w_1, w_2, \ldots, w_n$, where these words contain one *_blank_* and at least one entity, our goal is to fill in this *_blank_* with an answer entity q_a using a knowledge base $\mathcal{K}$ and text $\mathcal{T}$. Few example question answer pairs are shown in Table 2.

Universal Schema Traditionally universal schema is used for relation extraction in the context of knowledge base population. Rows in the schema are formed by entity pairs (e.g. USA, NYC), and columns represent the relation between them. A relation can either be a KB relation, or it could be a pattern of text that exist between these two entities in a large corpus. The embeddings of entities and relation types are learned by low-rank matrix factorization techniques. Riedel et al. (2013) treat textual patterns as static symbols, whereas recent work by Verga et al. (2016) replaces them with distributed representation of sentences obtained by a RNN. Using distributed representation allows reasoning on sentences that are similar in meaning but different on the surface form. We too use this variant to encode our textual relations.

Memory Networks MemNNs are neural attention models with external and differentiable memory. MemNNs decouple the memory component from the network thereby allowing it store external information. Previously, these have been successfully applied to question answering on KB where the memory is filled with distributed representation of KB triples (Bordes et al., 2015), or for read-

ing comprehension (Sukhbaatar et al., 2015; Hill et al., 2016), where the memory consists of distributed representation of sentences in the comprehension. Recently, key-value MemNN are introduced (Miller et al., 2016) where each memory slot consists of a key and value. The attention weight is computed only by comparing the question with the key memory, whereas the value is used to compute the contextual representation to predict the answer. We use this variant of MemNN for our model. Miller et al. (2016), in their experiments, store either KB triples or sentences as memories but they do not explicitly model multiple memories containing distinct data sources like we do.

3 Model

Our model is a MemNN with universal schema as its memory. Figure 1 shows the model architecture.

Memory: Our memory $\mathcal{M}$ comprise of both KB and textual triples from universal schema. Each memory cell is in the form of key-value pair. Let $(\mathrm{s},\mathrm{r},\mathrm{o}) \in \mathcal{K}$ represent a KB triple. We represent this fact with distributed key $\mathbf{k} \in \mathbb{R}^{2d}$ formed by concatenating the embeddings $\mathbf{s} \in \mathbb{R}^d$ and $\mathbf{r} \in \mathbb{R}^d$ of subject entity s and relation r respectively. The embedding $\mathbf{o} \in \mathbb{R}^d$ of object entity o is treated as its value $\mathbf{v}$.

Let $(\mathrm{s}, [\mathrm{w}_1, \ldots, \mathrm{arg}_1, \ldots, \mathrm{arg}_2, \mathrm{w}_n], \mathrm{o}) \in \mathcal{T}$ represent a textual fact, where arg_1 and arg_2 correspond to the positions of the entities 's' and 'o'. We represent the key as the sequence formed by replacing arg_1 with 's' and arg_2 with a special '_blank_' token, i.e., $k = [\mathrm{w}_1, \ldots, \mathrm{s}, \ldots, \text{\_blank\_}, \mathrm{w}_n]$ and value as just the entity 'o'. We convert k to a distributed representation using a bidirectional LSTM (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005), where $\mathbf{k} \in \mathbb{R}^{2d}$ is formed by concatenating the last states of forward and backward LSTM, i.e., $\mathbf{k} = \left[\overrightarrow{\mathrm{LSTM}}(\mathrm{k}); \overleftarrow{\mathrm{LSTM}}(\mathrm{k})\right]$. The value $\mathbf{v}$ is the embedding of the object entity o. Projecting both KB and textual facts to $\mathbb{R}^{2d}$ offers a unified view of the knowledge to reason upon. In Figure 1, each cell in the matrix represents a memory containing the distributed representation of its key and value.

Question Encoder: A bidirectional LSTM is also used to encode the input question q to a distributed representation $\mathbf{q} \in \mathbb{R}^{2d}$ similar to the key encoding step above.

Attention over cells: We compute attention weight of a memory cell by taking the dot product of its key $\mathbf{k}$ with a contextual vector $\mathbf{c}$ which encodes most important context in the current iteration. In the first iteration, the contextual vector is the question itself. We only consider the memory cells that contain at least one entity in the question. For example, for the input question in Figure 1, we only consider memory cells containing USA. Using the attention weights and values of memory cells, we compute the context vector $\mathbf{c_t}$ for the next iteration t as follows:

$$\mathbf{c_t} = \mathbf{W_t}\left(\mathbf{c_{t-1}} + \mathbf{W_p}\sum_{(k,v)\in\mathcal{M}}(\mathbf{c_{t-1}}\cdot\mathbf{k})\mathbf{v}\right)$$

where $\mathbf{c_0}$ is initialized with question embedding $\mathbf{q}$, $\mathbf{W_p}$ is a projection matrix, and $\mathbf{W_t}$ represents the weight matrix which considers the context in previous hop and the values in the current iteration based on their importance (attention weight). This multi-iterative context selection allows multi-hop reasoning without explicitly requiring a symbolic query representation.

Answer Entity Selection: The final contextual vector c_t is used to select the answer entity q_a (among all 1.8M entities in the dataset) which has the highest inner product with it.

4 Experiments

4.1 Evaluation Dataset

We use Freebase (Bollacker et al., 2008) as our KB, and ClueWeb (Gabrilovich et al., 2013) as our text source to build universal schema. For evaluation, literature offers two options: 1) datasets for text-based question answering tasks such as answer sentence selection and reading comprehension; and 2) datasets for KB question answering.

Although the text-based question answering datasets are large in size, e.g., SQuAD (Rajpurkar et al., 2016) has over 100k questions, answers to these are often not entities but rather sentences which are not the focus of our work. Moreover these texts may not contain Freebase entities at all, making these skewed heavily towards text. Coming to the alternative option, WebQuestions (Berant et al., 2013) is widely used for QA on Freebase. This dataset is curated such that all questions can be answered on Freebase alone. But since our goal is to explore the impact of universal schema, testing on a dataset completely answerable on a KB is not ideal. WikiMovies dataset (Miller et al., 2016) also has similar properties. Gardner and Krishnamurthy (2017) created a dataset with motivations similar to

Model	Dev. F_1	Test F_1
Bisk et al. (2016)	32.7	31.4
ONLYKB	39.1	38.5
ONLYTEXT	25.3	26.6
ENSEMBLE.	39.4	38.6
UNISCHEMA	**41.1**	**39.9**

Table 1: QA results on SPADES.

ours, however this is not publicly released during the submission time.

Instead, we use SPADES (Bisk et al., 2016) as our evaluation data which contains fill-in-the-blank cloze-styled questions created from ClueWeb. This dataset is ideal to test our hypothesis for following reasons: 1) it is large with 93K sentences and 1.8M entities; and 2) since these are collected from Web, most sentences are natural. A limitation of this dataset is that it contains only the sentences that have entities connected by at least one relation in Freebase, making it skewed towards Freebase as we will see (§ 4.4). We use the standard train, dev and test splits for our experiments. For text part of universal schema, we use the sentences present in the training set.

4.2 Models

We evaluate the following models to measure the impact of different knowledge sources for QA.

ONLYKB: In this model, MemNN memory contains only the facts from KB. For each KB triple (e_1, r, e_2), we have two memory slots, one for (e_1, r, e_2) and the other for its inverse (e_2, r^i, e_1).

ONLYTEXT: SPADES contains sentences with blanks. We replace the blank tokens with the answer entities to create textual facts from the training set. Using every pair of entities, we create a memory cell similar to as in universal schema.

ENSEMBLE This is an ensemble of the above two models. We use a linear model that combines the scores from, and use an ensemble to combine the evidences from individual models.

UNISCHEMA This is our main model with universal schema as its memory, i.e., it contains memory slots corresponding to both KB and textual facts.

4.3 Implementation Details

The dimensions of word, entity and relation embeddings, and LSTM states were set to d =50. The word and entity embeddings were initialized with word2vec (Mikolov et al., 2013) trained on 7.5 million ClueWeb sentences containing entities in Freebase subset of SPADES. The network weights were initialized using Xavier initialization (Glorot and Bengio, 2010). We considered up to a maximum of 5k KB facts and 2.5k textual facts for a question. We used Adam (Kingma and Ba, 2015) with the default hyperparameters (learning rate=1e-3, β_1=0.9, β_2=0.999, ε=1e-8) for optimization. To overcome exploding gradients, we restricted the magnitude of the ℓ_2 norm of the gradient to 5. The batch size during training was set to 32.

To train the UNISCHEMA model, we initialized the parameters from a trained ONLYKB model. We found that this is crucial in making the UNISCHEMA to work. Another caveat is the need to employ a trick similar to *batch normalization* (Ioffe and Szegedy, 2015). For each minibatch, we normalize the mean and variance of the textual facts and then scale and shift to match the mean and variance of the KB memory facts. Empirically, this stabilized the training and gave a boost in the final performance.

Question	Answer
1. USA have elected _blank_, our first african-american president.	Obama
2. Angelina has reportedly been threatening to leave _blank_.	Brad_Pitt
3. Spanish is more often a second and weaker language among many _blank_.	Latinos
4. _blank_ is the third largest city in the United_States.	Chicago
5. _blank_ was Belshazzar 's father.	Nabonidus

Table 2: A few questions on which ONLYKB fails to answer but UNISCHEMA succeeds.

4.4 Results and Discussions

Table 1 shows the main results on SPADES. UNISCHEMA outperforms all our models validating our hypothesis that exploiting universal schema for QA is better than using either KB or text alone. Despite SPADES creation process being friendly to Freebase, exploiting text still provides a significant improvement. Table 2 shows some of the questions which UNISCHEMA answered but ONLYKB failed. These can be broadly classified into (a) relations that are not expressed in Freebase (e.g., african-american presidents in sentence 1); (b) intentional facts since curated databases only represent concrete facts rather than intentions (e.g., threating to leave in sentence 2); (c) comparative predicates like first, second, largest, smallest (e.g., sentences 3 and 4); and (d) providing addi-

Model	Dev. F_1
ONLYKB correct	39.1
ONLYTEXT correct	25.3
UNISCHEMA correct	41.1
ONLYKB or ONLYTEXT got it correct	45.9
Both ONLYKB and ONLYTEXT got it correct	18.5
ONLYKB got it correct and ONLYTEXT did not	20.6
ONLYTEXT got it correct and ONLYKB did not	6.80
Both UNISCHEMA and ONLYKB got it correct	34.6
UNISCHEMA got it correct and ONLYKB did not	6.42
ONLYKB got it correct and UNISCHEMA did not	4.47
Both UNISCHEMA and ONLYTEXT got it correct	19.2
UNISCHEMA got it correct and ONLYTEXT did not	21.9
ONLYTEXT got it correct and UNISCHEMA did not	6.09

Table 3: Detailed results on SPADES.

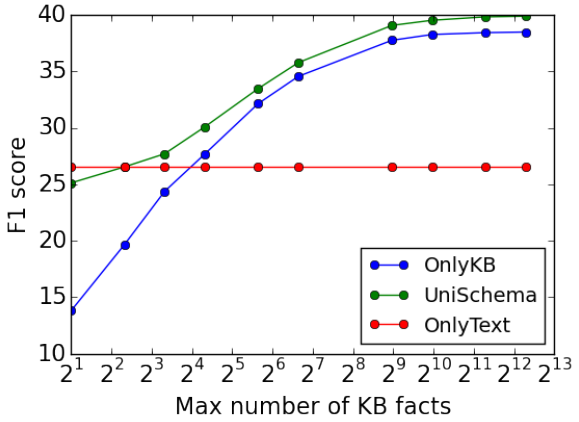

Figure 2: Performance on varying the number of available KB facts during test time. UNISCHEMA model consistently outperforms ONLYKB

tional type constraints (e.g., in sentence 5, Freebase does not have a special relation for *father*. It can be expressed using the relation *parent* along with the type constraint that the answer is of gender *male*).

We have also anlalyzed the nature of UNISCHEMA attention. In 58.7% of the cases the attention tends to prefer KB facts over text. This is as expected since KBs facts are concrete and accurate than text. In 34.8% of cases, the memory prefers to attend text even if the fact is already present in the KB. For the rest (6.5%), the memory distributes attention weight evenly, indicating for some questions, part of the evidence comes from text and part of it from KB. Table 3 gives a more detailed quantitative analysis of the three models in comparison with each other.

To see how reliable is UNISCHEMA, we gradually increased the coverage of KB by allowing only a fixed number of randomly chosen KB facts for each entity. As Figure 2 shows, when the KB coverage is less than 16 facts per entity, UNISCHEMA outperforms ONLYKB by a wide-margin indicating UNISCHEMA is robust even in resource-scarce scenario, whereas ONLYKB is very sensitive to the coverage. UNISCHEMA also outperforms ENSEMBLE showing joint modeling is superior to ensemble on the individual models. We also achieve the state-of-the-art with 8.5 F_1 points difference. Bisk et al. use graph matching techniques to convert natural language to Freebase queries whereas even without an explicit query representation, we outperform them.

5 Related Work

A majority of the QA literature that focused on exploiting KB and text either improves the inference on the KB using text based features (Krishnamurthy and Mitchell, 2012; Reddy et al., 2014; Joshi et al., 2014; Yao and Van Durme, 2014; Yih et al., 2015; Neelakantan et al., 2015b; Guu et al., 2015; Xu et al., 2016b; Choi et al., 2015; Savenkov and Agichtein, 2016) or improves the inference on text using KB (Sun et al., 2015).

Limited work exists on exploiting text and KB jointly for question answering. Gardner and Krishnamurthy (2017) is the closest to ours who generate a open-vocabulary logical form and rank candidate answers by how likely they occur with this logical form both in Freebase and text. Our models are trained on a weaker supervision signal without requiring the annotation of the logical forms.

A few QA methods infer on curated databases combined with OpenIE triples (Fader et al., 2014; Yahya et al., 2016; Xu et al., 2016a). Our work differs from them in two ways: 1) we do not need an explicit database query to retrieve the answers (Neelakantan et al., 2015a; Andreas et al., 2016); and 2) our text-based facts retain complete sentential context unlike the OpenIE triples (Banko et al., 2007; Carlson et al., 2010).

6 Conclusions

In this work, we showed universal schema is a promising knowledge source for QA than using KB or text alone. Our results conclude though KB is preferred over text when the KB contains the fact of interest, a large portion of queries still attend to text indicating the amalgam of both text and KB is superior than KB alone.

Acknowledgments

We sincerely thank Luke Vilnis for helpful insights. This work was supported in part by the Center for Intelligent Information Retrieval and in part by DARPA under agreement number FA8750-13-2-0020. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to Compose Neural Networks for Question Answering. In *NAACL*.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *IJCAI*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *EMNLP*.

Yonatan Bisk, Siva Reddy, John Blitzer, Julia Hockenmaier, and Mark Steedman. 2016. Evaluating Induced CCG Parsers on Grounded Semantic Parsing. In *EMNLP*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *ICDM*.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR* .

Razvan C. Bunescu and Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *ACL*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Jr. Estevam R. Hruschka, and Tom M. Mitchell. 2010. Toward an Architecture for Never-ending Language Learning. In *AAAI*.

Eunsol Choi, Daniel Hewlett, Alexandre Lacoste, Illia Polosukhin, Jakob Uszkoreit, and Jonathan Berant. 2016. Hierarchical question answering for long documents. *arXiv preprint arXiv:1611.01839* .

Eunsol Choi, Tom Kwiatkowski, and Luke Zettlemoyer. 2015. Scalable Semantic Parsing with Partial Ontologies. In *ACL*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. New York, NY, USA, KDD '14.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *KDD*. ACM, pages 1156–1165.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, and others. 2010. Building Watson: An overview of the DeepQA project. *AI magazine* .

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora. `(http://lemurproject.org/clueweb09/`.

Matt Gardner and Jayant Krishnamurthy. 2017. Open-Vocabulary Semantic Parsing with both Distributional Statistics and Formal Knowledge. In *AAAI*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* .

Bert F Green Jr, Alice K Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*. ACM, pages 219–224.

K. Guu, J. Miller, and P. Liang. 2015. Traversing knowledge graphs in vector space. In *EMNLP*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. *ICLR* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* .

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*. JMLR Workshop and Conference Proceedings.

Mandar Joshi, Uma Sawant, and Soumen Chakrabarti. 2014. Knowledge Graph and Corpus Driven Segmentation and Answer Inference for Telegraphic Entity-seeking Queries. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* .

Jayant Krishnamurthy and Tom Mitchell. 2012. Weakly Supervised Training of Semantic Parsers. In *EMNLP*.

Julian Kupiec. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *SIGIR*. ACM.

Kenton Lee, Tom Kwiatkowski, Ankur Parikh, and Dipanjan Das. 2016. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *In EMNLP* .

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Arvind Neelakantan, Quoc V Le, and Ilya Sutskever. 2015a. Neural programmer: Inducing latent programs with gradient descent. *arXiv preprint arXiv:1511.04834* .

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. 2015b. Compositional vector space models for knowledge base completion. In *ACL*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*. Austin, Texas.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *TACL* 2.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML PKDD*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *NAACL*.

Denis Savenkov and Eugene Agichtein. 2016. When a knowledge base is not enough: Question answering over knowledge bases with external text data. In *SIGIR*. ACM.

Minjoon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Query-reduction networks for question answering. *arXiv preprint arXiv:1606.04582* .

Gabriel Stanovsky, Omer Levy, and Ido Dagan. 2014. Proposition Knowledge Graphs. *COLING 2014* .

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*.

Huan Sun, Hao Ma, Wen-tau Yih, Chen-Tse Tsai, Jingjing Liu, and Ming-Wei Chang. 2015. Open domain question answering via semantic enrichment. In *WWW*. ACM.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A Machine Comprehension Dataset. *CoRR* abs/1611.09830.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema .

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*. volume 99, pages 77–82.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905* .

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. *In ICLR* .

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic Coattention Networks For Question Answering. *arXiv preprint arXiv:1611.01604* .

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016a. Hybrid Question Answering over Knowledge Base and Free Text. In *COLING*.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016b. Question Answering on Freebase via Relation Extraction and Textual Evidence. In *ACL*.

Mohamed Yahya, Denilson Barbosa, Klaus Berberich, Qiuyue Wang, and Gerhard Weikum. 2016. Relationship queries on extended knowledge graphs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pages 605–614.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *EMNLP*.

Xuchen Yao and Benjamin Van Durme. 2014. Information Extraction over Structured Data: Question Answering with Freebase. In *ACL*.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In *ACL*.

John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *AAAI*. Portland, Oregon.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *UAI*. Edinburgh, Scotland.