# A Review of Neural Approaches to the Question Answering Task

William Needham

City, University of London, London, EC1V 0HB, United Kingdom
william.needham@city.ac.uk

**Abstract.** The Question Answering task, whereby a system receives a plain language question from a user and returns a concise answer from a corpus of documents, has received considerable attention from academia and the commercial world since mid-way through the $20^{th}$ century. This paper offers a concise overview of this literature, focussing on recent advancements of the state-of-the-art achieved by neural network-based approaches. The rate of change of these advancements is considerable and has left a sparse landscape of analysis and research still to be conducted. My main contribution in this paper is to shine a light on these gaps in the literature, offering inspiration for future research.

**Keywords:** Information Retrieval · Question Answering · Neural Networks · Word embeddings · Pre-trained language models.



## 1 Introduction

A Question Answering (QA) system receives a human-language question, seeks to interpret large quantities of structured and unstructured data, and returns a concise answer (Hirschman and Gaizauskas, 2001). QA systems have been a vibrant field of research since the release of the Baseball system (Green Jr. et al., 1961). QA is an important task as typically users do not want to comprehend multiple, long documents to find an answer to their question (Lin et al., 2003). Throughout this time, we have seen a myriad of approaches, from knowledge-based approaches (Berant et al., 2013; Bollacker et al., 2008; Green Jr. et al., 1961) to information-retrieval based systems (Brill et al., 2002; Hirschman et al., 1999; Lin, 2007), as well as hybrids of the two such as the DeepQA system by IBM (Ferucci et al., 2010). Since the early 2000's, the introduction of neural-network-based approaches has resulted in marked success within the domain.

More recently, advances to the state-of-the-art have been attributed to the application of pre-trained language models; specifically the BERT (Devlin et al., 2018) language model (Bi-directional Encoder Representations from Transformers). On the SQuAD (Stanford Question Answering Dataset) benchmark

(Rajpurkar et al., 2018), every one of the top 20 submissions claims to have used a variation of BERT (as of April 2019). Just recently, human performance has been surpassed for the first time on this dataset.

It is clear the domain has progressed significantly within a short space of time. The speed of advancement has resulted in a shortage of published research explaining the architectures of such systems and perhaps more critically, understanding where these models are under-performing. This is important as without a clear understanding of the weaknesses of each implementation, it is difficult to improve the model with a subsequent iteration. This short paper seeks to identify these most prominent gaps in the literature, offering fruitful directions for future research.

The remainder of this paper is structured as follows: § 2 provides an overview of the traditional (non-neural) methods for solving the QA task. Then, in § 3, the core components of neural network architectures are described (specific to the IR/ QA task), followed by a comprehensive review of the neural architectures which combine these components in § 4, including extensions to the previously described traditional methods. § 5 offers an overview of the datasets and metrics used in the Question Answering task. Finally, a summary of the future research directions is presented in § 6.

## 2 Traditional approaches

*How many games did the Yankees play in July?* This question was asked of the BASEBALL system (Green Jr. et al., 1961), one of the earliest QA systems in the literature. Whilst ground-breaking in its approach, the paper was clear on its limitations; limitations that set the path for decades of subsequent research on QA systems.

Before a detailed look at neural approaches to the QA task, a reflection on the traditional approaches, that laid the foundations for current research, will be presented.

### 2.1 Knowledge-based approaches

BASEBALL (Green Jr. et al., 1961) can be described as a knowledge-based question answering (KB-QA) system in that it seeks to build a structured semantic representation of the question, upon which it can query a structured database to return an answer. For example, a knowledge-based system would seek to parse the input question "When did John F Kennedy die?" into a semantic query representation such as Death-Year("John F Kennedy", x), or similar, which can then be used to query a structured knowledge base. More recently, this general principle has evolved by focussing on the extraction of Resource Description Framework (RDF) triplets from large-scale internet corpora, and storing these in a knowledge base for querying later (for examples, see Bollacker et al., 2008; Fader et al., 2011; Lehmann et al., 2012). However, the task of encoding knowledge is expensive and time-consuming (Clark and Porter, 1999) and KB-QA systems typically fail on questions from unseen domains (Abujabal et al., 2018).

### 2.2 Information retrieval-based approaches

Information retrieval-based question answering (IR-QA) systems search large corpora of textual documents (for example, the web) for documents or passages relevant to the input question. Once a relevant document has been retrieved by the IR-QA system, reading comprehension algorithms are applied to understand the text and return the most relevant answer (Jurafsky and Martin, 2008). Put simply, IR-QA systems search raw text (extracted keywords, for example), whereas KB-QA systems search knowledge bases (Park et al., 2014). An advantage of IR-QA systems over KB-QA systems is that knowledge does not have to be encoded prior to search, however their performance is somewhat dependant of the competence of the IR algorithm.

### 2.3 Statistical machine learning-based approaches

Machine learning-based (ML) approaches have garnered more success than the preceding rule-based methods and have been applied to both the question classification (Metzler and Croft, 2005; Nguyen et al., 2007) and answer selection (Suzuki et al., 2002) sub-problems. One drawback, however, is that they typically require hand-crafted feature engineering in collaboration with a domain expert.

## 3 Building blocks of neural architectures

Having covered traditional methods in the previous section, this section will explore neural architectures which perform the QA task in new ways. The section begins with an overview of the core components in a neural network system designed for the question answering task.

### 3.1 Word embeddings

In contemporary literature, neural networks have been successfully applied to every part of the QA system. The introduction of word embeddings (Mikolov et al., 2013) brought the worlds of neural network research and natural language processing closer together. This approach seeks to develop distributed representations of words and phrases as numeric vectors, which allows them to be trained using neural networks.

Whilst word2Vec (Mikolov et al., 2013) and the subsequent GloVe embeddings (Pennington et al., 2014) were successful for a variety of natural language tasks, researchers began to understand their limitations. Peters et al. (2018) noticed that the meaning of a word very much depends on the context in which it is written. For example, in the two sentences (1) '*Let's stick to improvisation in this skit*', and (2) '*The dog walker threw the stick far away*' the word '*stick*' has different meanings. In respect of this, Peters et al. proposed ELMo for deep contextual word embeddings proposed by Peters et al (ELMo) which, when applied

to existing NLP models, outperformed the state-of-the-art results for every task it was tested on, including the Stanford Question Answering dataset (Rajpurkar et al., 2018).

### 3.2 Convolutions

Given a row vector, or matrix, a convolution is a sliding window (or filter or kernel) applied across the input vector to produce an output. In a convolutional neural network, the optimal values of the kernel are learnt from labelled input data. Convolutional neural networks have proved extremely successful within computer vision. In a natural language setting, a sliding window (kernel) is passed over some predefined number of words. Perhaps the most common use of convolutions for textual data is character-level convolutional neural networks (Char-CNNs) introduced by Zhang et al. (2015).

### 3.3 Attention

An attention mechanism in a standard encoder/decoder network allows the decoder to look back at the hidden states from the input sequence, presented to the decoder as a new input of weighted averages (Bahdanau et al., 2015). Since its introduction, attention has received considerable research attention from the field (Britz et al., 2017; Luong et al., 2015; Vaswani et al., 2017). Using the weighted average of some hidden state is not only limited to the input sequence. In self-attention (Cheng et al., 2016), relations between different positions of the same inputs sequence are introduced. Vaswani et al. (2017) took this approach one step further by introducing an architecture based solely on self-attention, with no convolutions or recurrent properties. This foundations provide the building blocks for general language models, such as BERT.

## 4 Neural architectures for the QA task

The following sub-section will investigate specific neural architectures which make use of the above components specifically for the QA task.

### 4.1 Neural extensions to KB-QA

The KB-QA concepts above have been extended to include aspects of neural network architectures. Yin et al. (2016) propose a novel entity linking and ranking method for relatively simple factoid question answering from Freebase (Bollacker et al., 2008). Then, neural networks are used to (1) match between questions and fact candidates using a character-level Convolutional Neural Network (char-CNN), and (2) match between Freebase predicate and question's pattern using a word-level CNN (word-CNN).

Also using Freebase, Dong et al. (2015) posit that to advance beyond simple factoid QA, distributed representations of answer path, answer context, and

answer type must be learnt. To do this, they propose multi-column convolutional neural networks (MCCNNs) which learn these representations from question-answer pairs.

Finally, instead of semantic parsing to a vector representation, Sorokin and Gurevych (2018) proposed a graph representation instead; this then enables the use of Gated Graph Neural Networks (GGNNs) for the QA task. This approach resulted in a 27.4% improvement (F1 score) over the best non-graph-based model. GGNNs were first proposed by Li et al. (2016) for sequential modelling problems and extend the original GNN framework whereby a neural network receives a graph as input, performs a computation over the nodes and edges and returns a graph as output.

### 4.2 Neural extensions to IR-QA

Important work by Burges et al. (2005) introduced RankNet; a pairwise method for optimising a ranking of a list according to a traditional IR metric (such as Mean Reciprocal Rank) using gradient descent. Technically, 'RankNet' can be any model for which the output is a differentiable function, such as neural networks or even boosted trees. Since publishing RankNet, the authors have developed the idea further with LambdaRank and then LambdaMART. A summary of each is available in Burges (2010). Researchers from IBM combined this approach with Supervised Kemeny aggregation in Agarwal et al. (2012).

Practical implementation of Learning-To-Rank is now widely available through the TF-Ranking TensorFlow package (Pasumarthi et al., 2018).

### 4.3 Language modelling

Language modelling is the task of predicting the probability of the next word in a sentence. Although the technique has developed away from the Question Answering task specifically, it is now a fundamental concept in many state-of-the-art approaches.

Before Bengio et al. (2003), back-off tri-gam models (Katz, 1987) and smoothed tri-gram models (Jelinek, and Mercer, 1980; Kneser and Ney, 1995) were favoured among the research community. In his paper, 'A Neural Probabilistic Language Model', Bengio described how neural networks can be applied to learn this distributed representation of words and kick-started a new direction for the field.

### 4.4 Recurrent neural networks

Since then, neural language modelling has developed from simple feed-forward networks, to recurrent neural networks (Mikolov et al., 2010) and Long Short-Term Memory architectures (Graves, 2013). Sequence-to-sequence models (Seq2Seq) were introduced by Sutskever et al. (2014) and featured an encoder/decoder architecture which successfully mapped input sequences of words/ tokens to an output sequence. Seq2Seq have been successfully applied to machine translation, natural language generation and QA tasks. One drawback of the Seq2Seq

model, however, is that between the encoder and decoder layers, information is compressed into a fixed-length 'thought' vector. For short phrases this is acceptable, but as length of input sequence increases, errors in the decoding step surface. To overcome this, the concept of attention was introduced by Bahdanau et al. (2015).

### 4.5 Pre-trained language models

The current state-of-the-art for QA systems is based on pre-trained models which combine many of the concepts discussed previously and were first proposed by Dai and Le (2015). Pre-trained models are trained over two distinct phases. Firstly, an unsupervised model is trained on a very large open-domain corpus; Wikipedia in the case of the Bidirectional Encoder Representations from Transformer models, known as BERT (Devlin et al., 2018), and WebText corpus in the case of Generalised Pre-Trained models, known as GPT (Radford et al., 2019, 2018). Then this general-purpose language model can be fine-tuned to a downstream task (like Question Answering) by means of a small-scale supervised learning phase (Ramachandran et al., 2017).

This approach has been hugely successful, especially when applied to the QA task. For example, for the the widely used SQuAD 2.0 benchmark for QA systems (Rajpurkar et al., 2018), every one of the top 20 models on the leaderboard is some variant on the BERT model (Devlin et al., 2018).

## 5 Task

This section will explore the published datasets and benchmarks used to compare and evaluate QA models.

### 5.1 Datasets

As QA systems become more competent, the benchmarks used to assess them also need to change. Modern QA systems have surpassed the level required by some of the earlier benchmarks, including WikiQA (Yang et al., 2015), NewsQA (Trischler et al., 2017) and SQuAD 1.0 (Rajpurkar et al., 2016). In response, the community has proposed new datasets which demand more capable systems. The bABi story dataset (Weston et al., 2015) requires logical reasoning, the SQuAD 2.0 dataset (Rajpurkar et al., 2018) includes unanswerable questions, and the CODAH dataset (Chen et al., 2019) introduces adversarial questioning. All of these datasets, however, were created through an artificial crowdsourcing methodology, which some have criticised as not being representative of the kind of questions humans would ask. The Natural Questions (NQ) dataset was recently released by Google (Kwiatkowski et al., 2019) in response to this criticism. The Natural Questions dataset consists of 307,372 training examples, 7,830 development examples and 7,842 examples in a hidden test set. For each question, both a long answer (span) and a short answer are expected.

### 5.2 Evaluation metrics

Evaluation metrics for the QA task vary based on the benchmark being used. WikiQA reported results for both MAP (Mean Averaged Precision) and MRR (Mean Reciprocal Rank). Alternatively, the SQuAD benchmark uses both Exact Match (EM) and macro-averaged F1 to assess submissions. NewsQA also uses EM and the F1 score, and also evaluates the BLEU score (Papineni et al., 2002) and CIDEr score (Vedantam et al., 2015). The more recent Natural Questions dataset reports Precision, Recall and the F1 score. This diversity of metrics reflects the diversity of approaches, with researchers coming from both information retrieval and machine learning backgrounds to tackle the QA task using metrics they are familiar with from their respective fields.

## 6 Future research directions

Whilst it is encouraging to see consistent advancement of the state-of-the-art, the rate of advancement has left considerable gaps in the research landscape. A summary of the exposed gaps in the existing literature is presented:

1. Is the model fully understanding the question? Mudrakarta et al. (2018) have begun to explore this direction, but more research is definitely required.
2. Is SQuAD (Rajpurkar et al., 2018) a suitable benchmark for the QA task? Some criticism has been raised over the synthetic nature in which the questions are produced (Kwiatkowski et al., 2019). Models need to be analysed against new benchmarks, such as Kwiatkowski et al.'s Natural Questions dataset and the CODAH benchmark (Chen et al., 2019).
3. The existing literature is lacking in an analysis on ensemble methods applied to the QA task. How can we quantitatively and qualitatively reason about why an ensemble of two models may or may not be effective, based on their individual strengths and weaknesses?
4. A degradation in performance is typical when assessing QA systems on datasets that require an element of logical reasoning. Schlag and Schmidhuber (2018) have began research in this direction but further research is required.
5. Another powerful pre-trained language model has recently been released by OpenAI - GPT2 (Radford et al., 2019). This model shows promise but it is yet to be properly analysed by the research community. OpenAI highlighted lack of 'world-knowledge' as a limitation of the GPT2 model and this is another avenue to be explored further.

## References

1. Abujabal, A., Roy, R.S., Yahya, M., Weikum, G., 2018. Never-Ending Learning for Open-Domain Question Answering over Knowledge Bases. WWW 18 Proc. 2018 World Wide Web Conf. 1053–1062.

2. Agarwal, A., Raghavan, H., Subbian, K., Melville, P., Lawrence, R.D., Gondek, D.C., Fan, J., 2012. Learning to Rank for Robust Question Answering, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12. ACM, New York, NY, USA, pp. 833–842. https://doi.org/10.1145/2396761.2396867
3. Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR.
4. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A Neural Probabilistic Language Model. J. Mach. Learn. Res. 3, 1137–1155.
5. Berant, J., Chou, A., Frostig, R., Liang, P., 2013. Semantic Parsing on Freebase from Question-Answer Pairs. Proc. 2013 Conf. Empir. Methods Nat. Lang. Process. 1533–1544.
6. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. SIGMOD 08 Proc. 2008 ACM SIGMOD Int. Conf. Manag. Data 1247–1250.
7. Brill, E., Dumais, S., Banko, M., 2002. An Analysis of the AskMSR Question-Answering System. Proc. Conf. Empir. Methods Nat. Lang. Process. EMNLP 257–264.
8. Britz, D., Goldie, A., Luong, M.-T., Le, Q., 2017. Massive Exploration of Neural Machine Translation Architectures, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp. 1442–1451. https://doi.org/10.18653/v1/D17-1151
9. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G., 2005. Learning to rank using gradient descent, in: Proceedings of the 22nd International Conference on Machine Learning - ICML '05. pp. 89–96. https://doi.org/10.1145/1102351.1102363
10. Burges, C., 2010. From RankNet to LambdaRank to LambdaMART: An Overview (No. MSR-TR-2010-82).
11. Chen, D., Fisch, A., Weston, J., Bordes, A., 2017. Reading Wikipedia to Answer Open-Domain Questions. Presented at the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1870–1879.
12. Chen, M., D'Arcy, M., Liu, A., Fernandez, J., Downey, D., 2019. CODAH: An Adversarially Authored Question-Answer Dataset for Common Sense [WWW Document].
13. Cheng, J., Dong, L., Lapata, M., 2016. Long Short-Term Memory-Networks for Machine Reading, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp. 551–561.
14. Clark, P., Porter, B., 1999. A Knowledge-Based Approach to Question-Answering. AAAI'99 Fall Symp. Quest. Answering Syst. 43–51.
15. Dai, A.M., Le, Q.V., 2015. Semi-supervised Sequence Learning, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 3079–3087.
16. Devlin, J., Wang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
17. Dong, L., Wei, F., Zhou, M., Xu, K., 2015. Question Answering over Freebase with Multi-Column Convolutional Neural Networks, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. pp. 260–269.

18. Fader, A., Soderland, S., Etzioni, O., 2011. Identifying Relations for Open Information Extraction. Proc. 2011 Conf. Empir. Methods Nat. Lang. Process. 1535–1545.
19. Ferucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, W., Nyberg, E., Prager, J., 2010. Building Watson: An Overview of the DeepQA Project. Assoc. Adv. Artif. Intell. 59–79.
20. Graves, A., 2013. Generating Sequences With Recurrent Neural Networks.
21. Green Jr., B.F., Wolf, A.K., Chomsky, C., Laughery, K., 1961. Baseball: an automatic question-answerer. Proceeding IRE-AIEE-ACM 61 West. 219–224.
22. Hirschman, L.A., Gaizauskas, R.J., 2001. Natural language question answering: the view from here. Nat. Lang. Eng. 7, 275–300.
23. Hirschman, L.A., Light, M., Breck, E., Burger, J., 1999. Deep Read: A Reading Comprehension System. Proc. 37th Annu. Meet. Assoc. Comput. Linguist. Comput. Linguist. 325–332.
24. Huang, X., Xhang, J., Li, D., Peng, L., 2019. Knowledge Graph Embedding Based Question Answering. WSDM 19 Proc. Twelfth ACM Int. Conf. Web Search Data Min. 105–113.
25. Jelinek, F., Mercer, R., 1980. Interpolated estimation of Markov source parameters from sparse data, in: Proceedings of the Workshop on Pattern Recognition in Practice.
26. Jurafsky, D., Martin, J., 2008. Speech and Language Processing, 2nd ed. Pearson International.
27. Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoust. Speech Signal Process. 35, 400–401.
28. Kneser, R., Ney, H., 1995. Improved backing-off for M-gram language modeling. Presented at the International Conference on Acoustics, Speech, and Signal Processing, IEEE.
29. Kwiatkowski, T., Palomaki, J., Redfield, O., 2019. Natural Questions: a Benchmark for Question Answering Research. Trans. Assoc. Comput. Linguist.
30. Lehmann, J., Furche, T., Grasso, G., 2012. deqa: Deep Web Extraction for Question Answering. Int. Semantic Web Conf. 2012 131–147.
31. Lin, J., 2007. Is Question Answering Better than Information Retrieval? Towards a Task-Based Evaluation Framework for Question Series. Hum. Lang. Technol. 2007 Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf. 212–219.
32. Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R.. Gated graph sequence neural networks. In ICLR, 2016.
33. Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., Karger, D., 2003. What Makes a Good Answer? The Role of Context in Question Answering. Proc. Ninth IFIP TC13 Int. Conf. Hum.-Comput. Interact. INTERACT 2003 25–32.
34. Luong, T., Pham, H., Manning, C.D., 2015. Effective Approaches to Attention-based Neural Machine Translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp. 1412–1421.
35. Metzler, D., Croft, W.B., 2005. Analysis of Statistical Question Classification for Fact-Based Questions. Inf Retrieval 8, 481–504.
36. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S., 2010. Recurrent Neural Network Based Language Model. Presented at the 1th Annual Conference of the International Speech Communication Association, pp. 1045–1048.

37. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. ArXiv13104546 Cs Stat.
38. Mudrakarta, P.K., Taly, A., Sundararajan, M., Dhamdhere, K., 2018. Did the Model Understand the Question?, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 1896–1906.
39. M.L. Nguyen, T.T. Nguyen and A. Shimazu. 2007. Subtree Mining for Question Classification Problem. In Proc. of IJCAI 2007
40. Otsuka, A., Nishida, K., Bessho, K., Asano, H., Tomita, J., 2018. Query Expansion with Neural Question-to-Answer Translation for FAQ-based Question Answering, in: Companion Proceedings of the The Web Conference 2018, WWW '18. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1063–1068.
41. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318.
42. Park, S., Shim, H., Lee, G.G., 2014. ISOFT at QALD-4: Semantic Similarity-based Question Answering System over Linked Data. CLEF CEUR Workshop.
43. Pasumarthi, R.K., Bruch, S., Wang, X., Li, C., Bendersky, M., Najork, M., Pfeifer, J., Golbandi, N., Anil, R., Wolf, S., 2018. TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. Presented at the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.
44. Pennington, J., Socher, R., Manning, C., 2014. Glove: Global Vectors for Word Representation. Presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.
45. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep Contextualized Word Representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.
46. Radford, A., Narasimhan, K., Salimans, T., Sutskever, T., 2018. Improving Language Understanding by Generative Pre-Training.
47. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners.
48. Rajpurkar, P., Jia, R., Liang, P., 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. Proc. 56th Annu. Meet. Assoc. Comput. Linguist. 2, 784–789.
49. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp. 2383–2392.
50. Ramachandran, P., Liu, P., Le, Q., 2017. Unsupervised Pretraining for Sequence to Sequence Learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp. 383–391.

51. Schlag, I., Schmidhuber, J., 2018. Learning to Reason with Third Order Tensor Products, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 9981–9993.
52. Sorokin, D., Gurevych, I., 2018. Modeling Semantics with Gated Graph Neural Networks for Knowledge Base Question Answering, in: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3306–3317.
53. Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to Sequence Learning with Neural Networks, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 3104–3112.
54. Suzuki, J., Sasaki, Y., Maeda, E., 2002. SVM Answer Selection for Open-Domain Question Answering, in: COLING 2002: The 19th International Conference on Computational Linguistics.
55. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K., 2017. NewsQA: A Machine Comprehension Dataset. Presented at the Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 191–200.
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5998–6008.
57. Vedantam, R., Zitnick, C.L., Parikh, D., 2015. CIDEr: Consensus-based image description evaluation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 4566–4575.
58. Weston, J., Bordes, A., Chopra, S., Rush, A.M., van Merriënboer, B., Joulin, A., Mikolov, T., 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks.
59. Yang, Y., Yih, W., Meek, C., 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp.
60. W. Yin, M. Yu, B. Xiang, B. Zhou, and H. Schutze. Simple question answering by attentive convolutional neural network. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers.
61. Zhang, X., Zhao, J., LeCun, Y., 2015. Character-level Convolutional Networks for Text Classification, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 649–657.